UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS MATEMÁTICAS
Departamento de Estadistica e Investigacion Operativa

TESIS DOCTORAL

Nuevas aportaciones a la morfosistemática estadística : estudio comparativo de métricas y su incidencia sobre la aplicación de métodos en taxonomía matemática

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR

Andrés María Gutiérrez Gómez

Madrid, 2015

Andrés María Gutierrez Gómez

NUEVAS APORTACIONES A LA MORFOSISTEMICA ESTADISTICA: ESTUDIO COMPARATIVO DE METRICAS Y SU INCIDENCIA SOBRE LA APLICACION DE METODOS EN TAXONOMIA MATEMATICA

Departamento de Estadística e Investigación Operativa
Facultad de Ciencias Matemáticas
Universidad Complutense de Madrid
1984

Colección Tesis Doctorales. Nº 29/84

Edita e imprime la Editorial de la Universidad
Complutense de Madrid. Servicio de Reprografía
Noviciado, 3 Madrid-8
Madrid, 1984
Xerox 9200 XB 480
Depósito Legal: M-3295-1984

NUEVAS APORTACIONES A LA MORFOSISTEMICA ESTADISTICA: "ESTUDIO COMPARATIVO DE METRICAS Y SU INCIDENCIA SOBRE LA APLICACION DE METODOS EN TAXONOMIA MATEMATICA"

Tesis Doctoral presentada por ANDRES Mª GUTIERREZ GOMEZ
en la Facultad de Ciencias Matemáticas de la Universidad Complutense

---

Realizada bajo la dirección del Prof. Dr. D. FRANCISCO AZORIN POCH,
y de la que es ponente el Prof. Dr. D. SIXTO RIOS GARCIA.
Madrid, junio 1982

A mis padres, Mª Carmen y Valeriano,
a mis hijos, Andrés, Javier, Ignacio y Blanca,
a mi esposa Blanca, verdadero motor y alma del presente trabajo.

El sacrificio y la ilusión de todos ellos han sido mi estímulo.

Es de justicia, expresar mi reconocimiento
al PROFESOR DOCTOR D. FRANCISCO AZORIN POCH,

bajo cuya dirección se ha llevado a cabo este trabajo; sus múltiples consejos, orientaciones, atenciones y la revisión del mismo, le hacen merecedor de mi gratitud.

Ayuda de inestimable valor ha sido la que me han prestado las señoritas Encarnación Navarro y Virginia Galera, que con ilusión e interés han copiado mis manuscritos.

PROLOGO

El presente trabajo, con el que aspiramos al grado de Doctor, se desarrolla a lo largo de seis capítulos y consta de dos partes bien diferenciadas.

Una primera parte, que está constituída por los cuatro primeros capítulos, recoge, sistematiza y ordena los principios básicos de la taxonomía numérica y del análisis de conglomerados así como los procedimientos existentes para la evaluación y comparación de las técnicas de dicho análisis.

Entendemos que esta primera parte era necesaria por dos razones: la primera, porque sirve de fundamento e introducción al resto del trabajo y, la segunda razón es que, a pesar del gran número de trabajos existentes, era necesaria una ordenación y sistematización de los conceptos aquí recogidos ya que se encontraban desperdigados.

Otras aportaciones nuestras en estos cuatro capítulos, son un catálogo de medidas de similitud y disimilitud que ofrecemos y que hemos subdividido en medidas de asociación, coeficientes angulares, medidas de distancia, medidas probabilísticas y medidas funcionales de similitud, nuevas medidas de distancia entre OTUS y entre conglomerados y un nuevo método de conglomeración que hemos denominado δ.

A la segunda parte, pertenecen los dos capítulos restantes.

En el capítulo V se ha clasificado, por primera vez, una colección de ejemplares del género NEUROPTERIS, PTERIDOSPERMEAS, fósiles del Carbonifero Superior, mediante cuatro métodos de conglomeración, distancia mínima, distancia máxima, método de la media y método δ , habiendo obtenido una clasificación objetiva que era muy necesaria en Paleobotánica.

Así mismo hemos comparado los resultados obtenidos por dichos métodos, mediante el coeficiente cofenético de SOKAL, comprobando la eficiencia y buenos resultados de nuestro método δ .

En el capítulo VI hemos vuelto el análisis de conglomerados sobre sí mismo para realizar un estudio comparativo de métricas. Se seleccionaron las métricas Euclidea, de manzanas, de Chebyshev, y de Canberra y, una vez definida en dicho conjunto la métrica δ^*, hemos conglomerado los resultados utilizando tres métodos: enlace sencillo, enlace completo y distancia promedio no ponderado.

Finalmente, y a partir de los correspondientes dendrogramas, hemos obtenido las matrices de valores cofenéticos que se han comparado mediante el coeficiente de correlación cofenético.

Es de observar que todo lo realizado en este último capítulo se ha hecho a partir de dos conjuntos totalmente distintos de OTUS, los elementos no metales del sistema periódico y los ejemplares de Neuropteris del capítulo V.

Entendemos que quedan abiertos caminos de investigación y que, con la ayuda de los medios informaticos se podrían repetir estos estudios ampliando el número de métricas a comparar y evaluar la sensibilidad de los resultados a las métricas y procedimientos utilizados.

INDICE GENERAL

CAPITULO III

RELACIONES ENTRE CONGLOMERADOS

CAPITULO I

IDEAS BASICAS DE TAXONOMIA NUMERICA Y ANALISIS DE CONGLOMERADOS

TAXONOMIA

TAXONOMIA NUMERICA

CLASIFICACION

I. OBJETIVOS:

II. MATERIAL:

III. CARACTERES Y SU MEDICION:

IV. ANALOGIA:

V. CRITERIOS DE CLASIFICACION:

VI. CLASES:

VII. PROCEDIMIENTOS Y ALGORITMOS: ANALISIS DE CONGLOMERADOS:

A.- EL PROBLEMA DE LA CONGLOMERACION

B.- DIVERSAS DEFINICIONES DE CONGLOMERADOS

C.- FUNDAMENTOS TEORICOS DEL ANALISIS DE CONGLOMERADOS

C.1. INTRODUCCION

C.2. FUNCIONES DE COMPARACION: SIMILITUDES Y DISIMILITUDES

D.- TECNICAS DE CONGLOMERACION

D.1. CARACTERISTICAS FUNDAMENTALES DE LOS GRANDES GRUPOS DE METODOS DE CONGLOMERACION.

D.2. CUALIDADES DE LOS METODOS DE CONGLOMERACION

E.- METODOS JERARQUICOS

F.- COMENTARIO SUCINTO DE LAS TECNICAS DE CONGLOMERACION MAS UTILIZADAS

F.1. ENLACE COMPLETO O DE LA DISTANCIA MAXIMA

F.2. ENLACE SENCILLO O DE LA MINIMA DISTANCIA

F.3. METODO DEL CENTROIDE

F.4. METODO DE LA MEDIANA

F.5. METODO DE LA DISTANCIA MEDIA

F.6. METODO DE CATTELL

F.7. CRITERIO DE OPTIMIZACION DE LA COVARIANZA

VIII. IDENTIFICACION Y ADSCRIPCION DE NUEVOS OBJETOS O UNIDADES TAXONOMICAS A CLASES YA ESTABLECIDAS:

UN NUEVO METODO DE CONGLOMERACION: EL METODO δ

TAXONOMIA

La TAXONOMIA (del griego TAXIS, arreglo, ordenación, y NOMOS, ley) es la ciencia de la clasificación y como dice GARRIDO MARECA"es instrumento para analizar la realidad, método de trabajo para agrupar observaciones y datos dispersos, clave de nuestro conocimiento de la naturaleza".

Adoptamos para el término TAXONOMIA, la siguiente definición. "El estudio de las bases, fundamentos, conceptos, métodos y problemas relacionados con la clasificación".

Los orígenes filosóficos del desarrollo actual de la taxonomía se remontan al trabajo del botánico francés del siglo XVIII Michel ADANSON (1760), aunque parece ser que el primero que utilizó la palabra taxonomía fue el botánico ginebrino A.P. de CANDOLLE (1778-1841).

Otro nombre que se ha dado a la taxonomía es el de SISTEMATICA. Toda ciencia implica sistematización, lo que quiere decir clasificación previa de sus ideas y de las materias en estudio.

A menudo, se utilizan indistintamente los términos TAXONOMIA, SISTEMATICA y CLASIFICACION. SIMPSON (115) nos da las siguientes definiciones para estos términos:

SISTEMATICA es el estudio científico de las clases y diversidad de organismos y de todas y cada una de las relaciones entre ellas.

CLASIFICACION es la ordenación de los organismos en grupos sobre la base de sus relaciones, es decir, a sus asociaciones por contigüidad, similitud o ambas.

Tal como está aquí definido el término clasificación nos indica un proceso, aunque se utiliza, a menudo, para designar el producto final del proceso.

TAXONOMIA es el estudio teórico de la clasificación, incluyendo sus bases, principios, procedimientos y reglas; al igual que ocurriría con el término anterior, la palabra taxonomía se ha usado para designar los productos finales del proceso taxonómico.

La taxonomía es una ciencia, y, como tal, no buscará la confección de catálogos sino que pretenderá lo universal, partiendo de una idea básica como es la consideración de los caracteres definitorios de los elementos a clasificar.

Afirma GARRIDO MARECA (49) que "la taxonomía es posible gracias a la existencia de discontinuidades. Existen diversas especies de seres y de objetos y, dentro de cada especie, pueden existir multitud de individuos todos ellos, con los caracteres de la especie a que pertenecen, pero diferenciándose por accidentes particulares, propios de los individuos o de los grupos de individuos que pueden constituir subespecies más o menos diferenciables".

Llegamos así a una cuestión importante, nudo gordiano al que se enfrentará cualquier taxónomo que pretenda formar conjuntos de seres u objetos homogéneos o comparables, y es: ¿qué objetos o seres englobará en un mismo grupo?; por supuesto, la contestación a la cuestión anterior estará íntimamente relacionada con la si--

guiente: ¿qué criterios debo utilizar? y aquí he personalizado a ese científico que investiga, puesto que no hemos de olvidar que la taxonomía, por su esencia, es una ciencia que, más que ninguna otra, está enormemente influenciada por las opiniones subjetivas de sus practicantes y en la medida en la que el taxónomo se independice de ese carácter subjetivo en la elección de "sus propiedades", así, saldrá una clasificación más o menos particular, más o menos científica y, creo, más o menos natural, puesto que al decir de SOKAL (21) "el propósito de la taxonomía es agrupar en clases NATURALES los objetos a clasificar".

La noción de CLASE NATURAL ha sido definida de distintas formas, pero detrás de todas ellas está la idea común de que un miembro cualquiera de una de estas clases naturales está más próximo a los de su misma clase que a cualquier otro.

Esto nos lleva a dos conceptos fundamentales: a) propiedades taxonómicas y b) relaciones taxonómicas.

En cuanto a las primeras, podríamos definirlas diciendo que "son aquellas propiedades inherentes a cualquier ser u objeto a clasificar que lo individualizan, independientemente de cualquier criterio de clasificación".

En cuanto a las relaciones taxonómicas, y, a pesar de que taxonomistas tradicionales pretenden identificarlas con las relaciones evolutivas, realmente, se pueden dividir en tres clases: FENETICAS, CLADISTICAS y CRONISTICAS.

Se llaman relaciones FENETICAS a aquellas relaciones basadas en una semejanza general entre los objetos a clasificar.

CLADISTICAS, a aquellas otras basadas en líneas comunes de descendencia y CRONISTICAS o TEMPORALES a las relaciones entre varias ramas de evolución.

Las clasificaciones FILOGENETICAS de la taxonomía convencional se basan, en general, en una mezcla indefinida de relaciones fenéticas y cladísticas y, en muchos casos, sólo representan una semejanza general entre los organismos clasificados, pero el término semejanza lo estudiaremos más tarde y ahora,no entramos en más consideraciones.

Terminamos este apartado, con la distinción que hizo GREGG (58) entre TAXONOMIA PROPIA y METODOLOGICA, según que se trate de observar, describir y clasificar, formulando y sometiendo a verificación ciertas leyes o bien, de resolver principios filosóficos generales relativos a la clasificación. Aplica la lógica simbólica y la teoría de conjuntos a la investigación de los fundamentos taxonómicos.

TAXONOMIA NUMERICA

El término TAXONOMIA NUMERICA y también el de TAXONOMIA MATEMATICA se utiliza en aquellos casos en los que se incluyen, no sólo el uso de mediciones, sino, también, la aplicación de otros conceptos matemáticos, como la partición de conjuntos discretos y continuos y los de analogía, proximidad o distancia, heterogeneidad, etc. (AZORIN POCH).

Según SNEATH y SOKAL (112), TAXONOMIA NUMERICA "es la evaluación, por métodos numéricos, de la afinidad o similitud entre unidades taxonómicas y el empleo de estas afinidades para construir un orden jerárquico de taxones".

U.H. HEYWOOD (66) dice que la TAXONOMIA NUMERICA es la evaluación de la semejanza entre grupos de organismos, y la ordenación de esos grupos en taxones de más alto rango, tomando como base estas semejanzas.

Cuando en la clasificación desempeñan papel principal las mediciones en sentido amplio, incluyendo determinación de modalidades o atributos, se suele emplear el término TAXIMETRIA o TAXOMETRIA que propuso GOOD, I.J. (también sugirió el nombre de BOTRIOMETRIA) para distinguirla de la taxonomía descriptiva o tradicional y de la clasificación aplicada a las ciencias naturales.

El desarrollo reciente de la taxonomía numérica se debe a los trabajos del microbiólogo británico SNEATH y de los entomólogos MICHENER y SOKAL.

La taxonomía numérica se basa en las ideas expresadas por primera vez por ADANSON, y que se describen concisamente en los siguientes axiomas modificados por SNEATH en 1958 (115):

I) La taxonomía ideal es aquella en la que los taxones tienen el mayor contenido de información y que se basa en la mayor cantidad posible de caracteres.

El término taxón es una abreviatura del grupo taxonómico de cualquier naturaleza y rango, según sugirieron LAM (1950) y RICKETT (1958).

II) A priori, cada carácter tiene igual peso en la creación de taxones naturales.

III) La similitud total o afinidad entre dos entidades, cualesquiera, es una función de la similitud de los muchos caracteres por los cuales se comparan.

IV) Taxones distintos se pueden construir por las correlaciones de diversos caracteres en los grupos bajo estudio.

V) La taxonomía se concibe, por tanto, como una ciencia estrictamente empírica.

VI) La AFINIDAD se estima independientemente de las consideraciones filogenéticas.

La taxonomía numérica que tiene como fines primordiales la repetibilidad y la objetividad, está basada en la evidencia, es decir, en las semejanzas presentadas por caracteres de taxo-

nes observados y registrados y no sobre probabilidades filogenéticas. No sólo es empírica, sino, también, operativa, es decir, emplea métodos que se basan sobre exposiciones e hipótesis formuladas de tal manera que pueden ser confirmadas por la observación y experimentación.

Las clasificaciones por medio de la taxonomía numérica se basan en una gran cantidad de caracteres expresados numéricamente.

Las operaciones lógicas que se realizan en la taxonomía numérica son las siguientes:

1.- Elección de las unidades que han de ser estudiadas: (objetos, individuos, razas, etc.).

Aquí se necesita el concepto de UNIDAD TAXONOMICA OPERACIONAL (OTU) que definieron SNEATH y SOKAL (1115) y que se estudiará en otro apartado.

2.- Selección del carácter: Los caracteres usados en taxonomía numérica tienen que ser descompuestos en caracteres UNITARIOS, que son aquellos que no pueden ser subdivididos en otros caracteres lógica o empíricamente independientes.

3.- Medida de la semejanza: La semejanza completa (S) se calcula por comparación de cada UTO con todos los demás y, normalmente, se expresa en porcentajes y se construye, entonces, una tabla de semejanza o matriz, tabulando los coeficientes (S) de cada UTO.

4.- Análisis de grupos: La matriz de semejanza se reordena ahora para que queden agrupadas aquellas UTO (S) cuyas componentes tienen la mayor semejanza entre sí. Esto puede realizarse por varios métodos y permite reconocer taxa o grupos relacionados que se denominan FENONES y que se pueden ordenar de manera jerárquica en un dendrograma.

5.- Discriminación: Una vez efectuada la clasificación, podemos volver sobre nuestros pasos y examinar de nuevo los caracteres para descubrir cuáles son los más constantes y, por tanto, los más válidos para construir claves y diagnosis.

Una última consideración obtenida de las clasificaciones numéricas es que no tienen que ser necesariamente jerárquicas en el sentido de formar grupos comprendidos, a su vez, en grupos mayores.

Los taxónomos numéricos proponen basar enteramente las clasificaciones sobre la semejanza, definiendo las clasificaciones naturales como aquellas cuyos miembros son en algún sentido más semejantes entre sí que a los miembros de otra clase.

De esta noción de "natural", basada en las ideas del botánico GILMOUR, se deduce que las clases naturales serán altamente predictivas, lo que se acerca a las ideas de ADANSON que era partidario de considerar al mayor número posible de caracteres, sin que predomine uno sobre los demás, lo que nos lleva a la semejanza fenética global, basándose así las clases naturales en el concepto esencialmente fenético de AFINIDAD.

Y terminamos este apartado, comentando dos aspectos fundamentales.

El primero está relacionado con la naturaleza de la semejanza, uno de los problemas fundamentales de la taxonomía que cuenta entre sus hipótesis fundamentales con la de que resulta posible cuantificar los grados de semejanza, que puede ser establecida sobre la base de caracteres homólogos o correspondientes. La homología, tal y como la interpretan los taxónomos, consiste más bien, en un semejanza global de la estructura que en una semejanza debida a una ascendencia común. 1'

El segundo aspecto está relacionado con el número de caracteres que deben elegirse para la descripción de la semejanza fenética.

Nos plantea SOKAL (21) dos alternativas que estimo interesantísimas:

¿Existe una semejanza asistótica entre los organismos s la que se va uno acercando a medida que aumenta el número de caracteres medidos? ¿o bien, cada conjunto adicional de caracteres contribuye a dotar a la semejanza de una nueva dimensión, haciendo, de este modo, que la estructura taxonómica de un grupo sea intrínsecamente inestable?

Esta importante cuestión aún no está dilucidada totalmente.

El propio SOKAL (21) afirma que si se quieren obtener medidas válidad de la semejanza global, se deben emplear tantos

conjuntos de caracteres, y tan variados, como sea posible:

"Si queremos establecer las clasificaciones atendiendo a la semejanza global, la taxonomía numérica deberá dar a sus procedimientos una base operacional y cuantitativa".

Algunos de estos procedimientos de la taxonomía numérica comenzaron a desarrollarse a principios de siglo, pero no dieron fruto hasta la aparición de las computadoras, debido, probablemente, a lo insuperable de las dificultades de cálculo.

CLASIFICACION

Según indica AZORIN POCH (5) en su trabajo "Estadística y taxonomía matemática", se pueden considerar los siguientes aspectos:

I. OBJETIVOS. Para qué se quiere clasificar.

II. MATERIAL. Qué se quiere clasificar.

III. CARACTERES Y SU MEDICION (puntuación, codificación, ponderación, etc.).

IV. ANALOGIA, PROXIMIDAD, HOMOGENEIDAD.

V. CRITERIOS DE CLASIFICACION.

VI. CLASES. Su definición, tipos de clases, agrupaciones o conglomerados, etc.

VII. PROCEDIMIENTOS Y ALGORITMOS: ANALISIS DE CONGLOMERADOS.

VIII. IDENTIFICACION Y ADSCRIPCION DE NUEVOS OBJETOS O UNIDADES A CLASES PREESTABLECIDAS.

Este último aspecto es el que algunos autores llaman, simplemente, CLASIFICACION. Corresponde al método estadístico denominado ANALISIS DISCRIMINANTE y constituye un caso especial del llamado RECONOCIMIENTO DE PATRONES.

Los anteriores puntos comprenden,especialmente, el análisis de agrupaciones o conglomerados y constituyen un caso particular de la construcción de patrones.

I. OBJETIVOS DE LA CLASIFICACION:

"En una divertida y clarificadora clasificación de clasificaciones, GOOD (23) señala cinco propósitos:

1. Para clarificación mental y comunicación.

2. Para descubrir nuevos campos de investigación.

3. Para planificación de una estructura orgánica.

4. Como lista de control.

5. Como diversión.

Para FLEISS y ZUBIN (23) el objetivo de una clasificación es llegar a una descripción útil de la muestra y descubrir conglomerados o clases no conocidos, que puedan ser importantes.

Según JARDINE (23), hay que clasificar,para representar los datos de tal modo que sugieran hipótesis fructíferas".

"Una clasificación es predictiva y su propósito preciso es desconocido en el momento de la clasificación. No puede ser cierta o incierta, probable o improbable, sólo útil o inútil (WILLIAMS y LANCE (1965)".

"Usualmente, una clasificación no se hace para obtener una respuesta a un problema planteado", JARDINE (23), aunque,a menudo se hace en la práctica, sin validez alguna, para probar una hipótesis, frecuentemente, de la existencia de los conglomerados que encuentra.

De todas formas, la primera decisión, con relación a los objetivos es medir hasta qué punto el esfuerzo y el tiempo compensan la clasificación.

Más concretamente, y siguiendo a AZORIN (7), habrá que decidir sobre los siguientes puntos:

a) Si conviene atender a uno o más objetivos, eligiendo entre mayor precisión o información o mayor sencillez y economía.

b) Si se quiere caracterizar en el sentido de que la pertenencia a una clase proporcione información sobre sus características, o sólamente "catalogar" o "etiquetar".

c) Si hay retroacción en el sistema, ya que pueden presentarse resultados que no se esperan, o dificultades de medición.

II. MATERIAL:

El material a clasificar puede ser discreto o continuo. A los elementos u objetos a clasificar se les denomina UNIDADES TAXONOMICAS OPERACIONALES, OTU. Pueden ser objetos en sentido co-

rriente, individuos, estirpes, especies, estímulos aplicados a sujetos, agrupaciones o clases con las que se van a formar otras más amplias, o abstracciones estadísticas de los grupos taxonómicos de orden más elevado.

Las clasificaciones, por medio de la taxonomía numérica, se hacen a partir de las unidades taxonómicas operacionales, basándonos en una gran cantidad de caracteres expresados numéricamente. Los caracteres son rasgos que definen una unidad; dentro de un determinado carácter, se pueden distinguir diversas modalidades.

A cada objeto se le asocia una lista de descripciones o vector de estados de carácter y la clasificación se lleva a cabo sobre una MATRIZ DE DATOS formada por una colección de vectores.

Los OTUS se representarán en un espacio cuyas dimensiones son los caracteres. Este espacio de atributos, A-espacio, es formalmente de n-dimensiones (para n caracteres) pero, a causa de las correlaciones de los caracteres de los OTUS, puede reducirse, generalmente, a un número menor de dimensiones, con una pequeña pérdida de información.

III. CARACTERES Y SU MEDICION:

El término "carácter" ha sido, al menos, empleado en dos formas distintas por los taxonomistas. Su uso más frecuente es como un rasgo que distingue a los taxones, una característica de una especie de organismos que lo distinguirá de cualquier otra. Este parece ser el sentido con que MAYR, LINSLEY y USINGER (115) definen un carácter taxonómico:

"Cualquier atributo de un organismo o de un grupo de organismos, por el que difiere de otro organismo perteneciente a una categoría taxonómica diferente o se asemeja a un organismo perteneciente a la misma categoría". Pero con estas definiciones, desafortunadas, caemos en un círculo vicioso, pues si el término carácter se restringe a las diferencias entre taxones, los taxones, por sí solos, no pueden ser reconocidos sin que los mismos caracteres sean conocidos primero.

Otro significado frecuente del término carácter, que ha sido expuesto por taxonomistas numéricos, es que un carácter es"un rasgo que varía de una especie de organismos a otra", MICHENER y SOKAL (112) o "si consideramos dos cosas al mismo tiempo, la que podamos tomar como variable independiente", CAIN y HARRISON (114).

Así, por ejemplo, si hablamos de figura de los cuerpos, se tratará de un carácter; y el ser redondo o cuadrado, serán diferentes "estados"[2*] del carácter, CAIN y HARRISON llamarían a éstos los "valores" de un carácter, pero según SNEATH y SOKAL es preferible el término primero, ya que éste no emplea una expresión cuantitativa y es, por lo tanto, más aceptable en casos de diferencias cualitativas. La palabra "estado" puede implicar una subdivisión cualitativa más que una subdivisión cuantitativa, pero en ausencia de un término más apropiado, lo empleamos en los dos sentidos (cualitativo y cuantitativo).

Así pues, podemos decir que los términos descriptivos aplicables a individuos o partes de individuos, son los estados de un carácter, por ejemplo, ser cuadrado.

Y el conjunto de términos descriptivos mutuamente exclu-

yentes será el carácter o propiedad, por ejemplo: figura de los cuerpos.

Un carácter puede ser CUALITATIVO, esto es lo que se llama en Estadística, un atributo, o bien, CUANTITATIVO o variable propiamente dicha, que toma valores numéricos.

Algunas cualidades pueden expresarse por medio de un número, resultado de una medida, y el atributo correspondiente, es un atributo cuantitativo continuo, tal como ocurre con el peso. Hay atributos cuyos estados no varían de un modo continuo, sino discreto, son atributos cuantitativos enumerables, tal como ocurre, por ejemplo, con la simetría; entre dos grupos de simetría no existe término medio. Los atributos no cuantificables ni enumerables se pueden expresar en términos de una dicotomía, bien por ausencia o presencia de un determinado carácter, o por una tricotomía: presencia de un carácter o su opuesto, o ausencia de ese carácter.

Un carácter será INTRINSECO cuando se refiera a propiedades propias de los individuos y, será EXTRINSECO si corresponde a relaciones de los individuos con seres o circunstancias de origen y filiación.

El número de caracteres que se pueden considerar en un individuo es muy grande, y, además debe indicarse el peso que se asignará a cada carácter, según la importancia que estimamos que tiene, aunque la idea de que algunos caracteres son más importantes que otros es muy vaga, algunas veces, lo que significa es que los estados de algunos caracteres nos dan un criterio de diagnóstico mejor que los de los otros.

El famoso botánico ADANSON (1760) era partidario de considerar el mayor número de caracteres y todos con igual peso, pero en la práctica, a veces, hay que reducir el número de los mismos, lo que equivale a atribuir ponderación cero a los caracteres excluídos. Por otra parte, tampoco se puede dar una justificación satisfactoria de las ponderaciones, ya que hay que tener presente que una vez establecidos los caracteres a considerar, los objetos o unidades en estudio quedan definidos por dichos caracteres, haciendo abstracción de todo lo demás.

Debe distinguirse, además, entre las ponderaciones y los coeficientes que se emplean en el cálculo de disimilitudes o distancias, para corregir las posibles correlaciones entre caracteres como ocurre en las expresiones de MAHALANOBIS y de IVANOVIC.

Por tanto, se tiene que, en el caso de variales, cada objeto queda definido por el vector de los valores de cada uno de sus caracteres o propiedades y puede representarse por un punto en el espacio euclídeo de tantas dimensiones y ejes coordenados como caracteres o variables.

Otro modo de representación es el perfil, que expresa los valores de los p-caracteres o propiedades de un objeto sobre perpendiculares en p-puntos sucesivos a distancias constantes en un mismo eje. Análogamente, se puede obtener el perfil de un carácter sobre perpendiculares en puntos correspondientes a los n individuos o unidades taxonómicas en estudio.

Los caracteres taxonómicos se pueden agrupar "a grosso modo" en:

1. Caracteres morfológicos (externos, internos, microscópicos, incluyendo caracteres citológicos y de desarrollo).
2. Caracteres funcionales.
3. Caracteres de comportamiento
4. Caracteres situacionales (ecológicos, de distribución habitat, alimentación, huéspedes, parásitos, dinámica de poblaciones, distribución geográfica).

Los caracteres pueden dividirse, también, en ESTATICOS (o FENETICOS) y DINAMICOS (o FILETICOS) según que se refieran a la situación o a la evolución del individuo o en SIMPLES Y COMPUESTOS.

Estas denominaciones se aplican, también, a las clasificaciones basadas en dichos caracteres o atributos. A veces, conviene agrupar los caracteres por facetas (GUTTMAN) o AFINIDADES.

Debemos distinguir, también, entre caracteres observados directamente y caracteres que resultan de aplicar a las observaciones una cierta transformación, siendo una de las más corrientes la estandarización para hacer comparables los valores de caracteres que se miden en unidades distintas.

Finalmente, los caracteres 3* se pueden dividir en INTRINSECOS, propios de los individuos, y EXTRINSECOS, correspondientes a relaciones de los individuos con seres o circunstancias exteriores a él o a circunstancias de origen y filiación.

Los atributos intrínsecos de los seres materiales pueden clasificarse de un modo natural en cuatro categorías:

Atributos GEOMETRICOS que se refieren a las relaciones de situación y medida y a las regularidades de los elementos que lo constituyen.

Atributos FISICOS, que son expresables por magnitudes referidas a sistemas de unidad.

Atributos QUIMICOS, que indican la constitución y la estructura atómica del ser o las variaciones de éstos frente a los agentes físicos o químicos.

Atributos BIOLOGICOS, que solo se presentan en los seres vivos y se refieren a las funciones de éstos, sus posibles relaciones con el medio y sus relaciones de origen y filiación.

El problema de encontrar medidas del parecido entre pares de entidades basadas en un determinado número de características, no es nuevo. Ha sido acometido en diversas ciencias siempre que era necesaria la labor clasificatoria.

Una vez elegidos los p-caracteres (variables) y obtenido el valor correspondiente de la medición para cada uno de ellos y en cada uno de los n objetos, individuos o unidades taxonómicas en estudio, se confecciona una matriz,llamada MATRIZ TAXONOMICA BASICA.

$$\begin{matrix} x_{11} & \cdots & \cdots & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & \cdots & x_{2p} \\ \vdots & & & & \vdots \\ x_{n1} & \cdots & \cdots & & x_{np} \end{matrix}$$

donde el elemento x_{ij} designa el valor numérico del j-ésimo carácter en el i-ésimo objeto.[4*]

Si se trata de atributos dicotómicos, X_{ij} se podría expresar por 1 (presencia del atributo) o 0 (ausencia de ese atributo, o presencia del contrario.

Cada vector o linea horizontal de la matriz taxonómica puede representarse por un solo símbolo correspondiente al vector fila individual p-caracterizado. Análogamente, cada vector columna puede representarse por un símbolo correspondiente a cada uno de los p vectores "caracteriales" n-individualizados.

A los vectores de individuos o "individuales" se les puede aplicar la llamada técnica Q y a los vectores de carácter o "caracterales", la técnica R.

Dada una matriz como la anterior, podemos examinarla desde dos puntos de vista. Si observamos la asociación de pares de unidades taxonómicas operacionales sobre todos los caracteres, estamos empleando la llamada técnica Q; y si hacemos lo contrario, es decir, la asociación de pares de caracteres sobre todas las unidades taxonómicas operacionales, es la llamada técnica R.

Las mencionadas técnicas Q y R han sido muy empleadas y en los más diversos campos desde la Psicología hasta la Ecología.

En taxonomía han sido empleadas ambas técnicas.En este campo, la técnica R se refiere a las correlaciones entre caracteres basadas en unidades taxonómicas operacionales. Estas podrían ir desde el rango más bajo posible (organismo individuales) hasta poblaciones, especies y taxones supraespecíficos. A un nivel infraespecífico se han llevado a cabo muchos trabajos del tipo R, entre los cuales podemos citar los de CLARK (1941), JOLICOEUR (1959), OLSON y MILLER (1958), SOKAL (1952, 1959, 1962), SOKAL y HUNTER (1955)

y SOKAL y RINKEL (1963), tales trabajos contribuyen principalmente a un entendimiento de procesos ontogénicos y de diferenciación genética secundaria. En los más altos niveles taxonómicos, el análisis de matrices del tipo R nos suministra información sobre factores filogenéticos.

Sin embargo, en taxonomía numérica son más importantes los estudios del tipo Q. Se refieren a las cuantificaciones de relaciones entre pares de taxones frecuentemente especies, basadas en un gran número de caracteres.

IV. ANALOGIA:

La analogía es el primer paso para la clasificación y se refiere a la participación de dos o más objetos en los valores o modalidades de sus caracteres.

Una vez que se han sustituído los objetos por las coordenadas que los definen, puede introducirse una expresión de analogía, proximidad, similitud o parecido, o bien, el concepto contrario, que si cumple las propiedades de identidad, simetría y triangularidad es una distancia propiamente dicha. Estas medidas de proximidad o alejamiento entre los puntos representativos de los objetos en estudio permiten construir una matriz Q de orden nxn, donde el elemento S_{ij} representa la similitud o disimilitud que existe entre el objeto i-ésimo y el objeto j-ésimo.

De manera análoga, puede construirse una matriz R de orden pxp. Así como la anterior daba las relaciones de similitud o disimilitud entre los $\binom{n}{2}$ pares de objetos, la R nos da dichas relaciones entre los $\binom{p}{2}$ pares de caracteres.

En el primer caso se obtiene una matriz simétrica

$$\begin{matrix} 1 & S_{12} & S_{12} \cdots \cdots & S_{1n} \\ S_{21} & 1 & \cdots \cdots \cdots \cdots & S_{2n} \\ S_{n1} & S_{n2} & \cdots \cdots \cdots \cdots & 1 \end{matrix}$$

en donde S_{ij} indica la similitud entre el objeto o unidad taxonómica i-ésima y la j-ésima.

Para todo i se verifica que $S_{i_i} = 1$

Análoagamente se obtendría una matriz con elementos d_{ij} que indican distancia o disimilitud:

$$\begin{matrix} 0 & d_{12} \cdots \cdots \cdots \cdots & d_{1n} \\ d_{21} & 0 \cdots \cdots \cdots \cdots & d_{2n} \\ d_{n1} & d_{n2} \cdots \cdots \cdots \cdots & 0 \end{matrix}$$

donde para todo i se verifica, $d_{ii} = 0$

Así pues, basadas en los mismos datos originales, obtendremos las matrices de correlación Q y R, donde Q representa correlaciones entre OTUS y R correlaciones entre caracteres.

V. CRITERIOS DE CLASIFICACION:

Distinguiremos, antes de estudiar los criterios, tres tipos de procedimientos de clasificación, que nos ofrece R.M. CORMACH (23).

1. CLASIFICACION JERARQUICA, en la que las clases se - clasifican ellas mismas en grupos, repitiéndose el proceso a diferentes niveles, hasta formar un árbol.

2. PARTICION, en la que las clases son mutuamente excluyentes, formando así una partición del conjunto de entidades a clasificar.

3. EN BLOQUES, en la que las clases o bloques pueden superponerse, considerándose como diferentes tipos de clases a un bloque y a su complemento.

También, debe hacerse la distinción entre las situaciones en las que a los elementos o entidades de una clase se les exija o no que sean "distantes" de los elementos de otra clase: situaciones denominadas por KENDALL (23), respectivamente, CLASIFICACION y DISECCION, "toda colección de elementos puede ser disecada, no todas pueden ser clasificadas".

"Si hay dos conglomerados densos de edificios separados por un gran espacio vacío, no existe dificultad en percibir la existencia de dos pueblos, mientras que si un pueblo con un nombre enlaza con otro de distinto nombre sentimos que la separación es artificial y que no existen dos entidades sino un GENGERELLI (52).

En lo que se refiere a los criterios de clasificación, que influyen más o menos según que los puntos sean dispersos y las fronteras borrosas o que sean claras las agrupaciones y más estructurado esté el patrón, podemos destacar las siguientes dicotomías:

1a) CLASIFICACIONES MONOTETICAS que se basan en la presencia o ausencia de los diferentes valores o modalidades de cada carácter, para la inclusión o exclusión en una clase.

1b) CLASIFICACIONES POLITETICAS, que consideran simultaneamente todos los caracteres establecidos, a fin de construir una matriz Q de similitudes o disimilitudes, como base de la clasificación.

La primera es más útil para la formación de claves de identificación o clasificación, y la segunda, para la constitución de agrupaciones, clases o conglomerados homogéneos.

2a) CLASIFICACIONES NITIDAS, de clases mutuamente excluyentes.

2b) CLASIFICACIONES BORROSAS,vagas o difusas, en las que un mismo elemento puede pertenecer a dos o más clases.

3a) CLASIFICACIONES NATURALES, en las que las clases están basadas en caracteres propios de su esencia.

3b) CLASIFICACIONES ARTIFICIALES, en las que los elementos se agrupan en clases establecidas de modo subjetivo. Se denominan, también, catálogos.

4a) CLASIFICACIONES JERARQUICAS, en las que se establecen clases o taxones a distintos niveles en diagrama

arborescente o dendrograma.

4b) CLASIFICACIONES NO JERARQUICAS, que solo presentan un nivel.

5a) CLASIFICACIONES FENETICAS, basadas en las similitudes o disimilitudes obtenidas a partir de los valores o modalidades de los objetos o unidades taxonómicas en los caracteres considerados.

5b) CLASIFICACIONES FILETICAS O CLADISTICAS, basadas en aspectos evolutivos y de ascendencia.

También, puede distinguirse entre clasificaciones con un número prefijado de clases y clasificaciones con grado prefijado de homogeneidad.

Al efectuar una clasificación ha de procurarse que la misma sea:

a. La más HOMOGENEA, es decir, aquella que esté constituída por conglomerados de elementos más parecidos entre sí, entre las que tengan el mismo número de conglomerados o clases.

b. La más ECONOMICA, ésto es, la que tiene menor número de clases, entre las que tienen un mismo grado de homogeneidad.

c. OPTIMA, o al menos aceptable para el par dual, homogeneidad y economía.

En el primer caso se empieza por fijar el número de conglomerados o clases que se considera adecuado, y se aplica un procedimiento de conglomeración hasta conseguir la más homogénea, de acuerdo con el criterio de homogeneidad que se establezca.

En el segundo caso, se empieza por fijar el grado de homogeneidad mínimo (mínimo admisible) y se aplica un procedimiento de conglomeración hasta alcanzar dicho grado con el menor número posible de clases.

En el tercer caso se parte, por ejemplo, de un número de clases provisional y se trata de ir ganando,en lo posible, en ambas características, mejor dicho, de ir ganando en una sin perder excesivamente en la otra.

Puede elegirse un número final de clases, tal que el aumentar en uno el número de éstas, no resulte importante o apreciable la ganancia en homogeneidad.

Según cita R. M. CORMACK (23) en su trabajo "A REVIEW OF CLASSIFICATION".

SILVESTRI y HILL (23) establecieron tres condiciones deseables para una clasificación biológica:

a'. OBJETIVIDAD. Investigadores independientes deben llegar a conclusiones similares.

b'. ESTABILIDAD. La clasificación debe ser poco afectada por la adición de nuevos datos.

c'. PREDICTIBILIDAD, de las variantes en nuevos individuos.

Estas condiciones no se aplican necesariamente a todos los campos. A veces, por ejemplo, en la clasificación de una colección completa de documentos de la cual ha de extraerse la información, no hay nuevos individuos y la c' no es aplicable.

"El requerimiento de estabilidad debe, también, tomarse en el sentido de implicación de robustez contra los errores en los datos", JONES y NEEDHAM (23).

Si una clasificación ha de permanecer virtualmente inalterable cuando se miden variables adicionales sobre los mismo objetos debe, claramente, ser capaz, en cierto sentido, de predecir variables no observadas aún, cosa que será posible si las variables no observadas están correlacionadas con las variables observadas.

Si una variable observada particular v es la variable única más altamente correlacionada con las otras variables observadas, es razonable suponer que, también es la más altamente correlacionada con las variables no observadas.

Esto conduce en ordenación, a la consideración de componentes principales, y en clasificación a métodos de estructura derivada, WILLIAMS y DALE (2), que no implican ninguna medida de "proximidad" entre individuos. Podremos dividir los individuos entre conglomerados de acuerdo con su valor de v.

VI. CLASES:

Partiendo de las unidades taxonómicas operacionales y

basándonos en los caracteres, se pueden establecer las CLASES o CONGLOMERADOS. Cada clase estará formada por un conjunto de UTO (S) con varios caracteres en común, lo cual no quiere decir que todos los elementos que estén en una misma clase, tengan el mismo número de caracteres en común.

Toda clasificación debe cumplir dos condiciones: a) que las categorías (ordenables o no) sean mutuamente excluyentes, y b) que sean exhaustivas.

Esto nos lleva a que las fronteras que delimitan dos clases distintas, deben estar bien definidas, aunque, a veces, hay que utilizar clasificaciones BORROSAS en las cuales las fronteras no sean excluyentes.

Es ventajoso poner los límites o fronteras en los puntos o zonas de rareza (mejor de ausencia, cuando sea posible) con el objeto de tener la máxima información.

Hemos de distinguir, también, entre agrupaciones o clases que tengan, sólamente, importancia en un estudio particular que hayamos realizado y clases que presentan una mayor permanencia. Surgen, así, tres conceptos fundamentales:

TAXON: Agrupación o partición establecida con carácter de permanencia, ampliamente reconocida y con nombre determinado. El taxón contiene una parte substancial de los caracteres considerados en una clasificación.
A una clase de este tipo no se llega a partir de una única propiedad, sino de un conjunto de ellas, y dos elementos, cualesquiera, de la misma clase, no tienen que tener nece-

sariamente los mismos caracteres comunes.

FENON: Es una agrupación o partición que puede completar o sustituir a los taxones, según los casos. Se le puede atribuir un nombre o número que exprese el grado de homogeneidad de las UTO (S) que lo constituyen.

CONGLOMERADO O CLASE: Agrupación o partición de cualquier naturaleza o jerarquía obtenida, como las anteriores, por métodos de taxonomía numérica y que se somete a estudio y análisis comparativo, con diversas finalidades, en particular, la de llegar a constituir fenones.

Según el diccionario de términos estadísticos, KENDALL y BUCKLAND definen el término conglomerado: como "un grupo de elementos contiguos de una población estadística". [5]

Otras definiciones son:

"Un conglomerado es un conjunto de entidades que son semejantes y entidades de diferentes conglomerados no lo son".

"Las entidades de un conglomerado son más similares unas a otras que las entidades de otros conglomerados".

WALLACE y BOULTON (23) proponen: "un conglomerado es un subconjunto de entidades que pueden, útilmente, ser tratadas como equivalentes en alguna discusión".

GENGRELLI (51) define un conglomerado como una agregación de puntos del espacio de prueba, tal que la DISTANCIA entre dos puntos, cualesquiera, del conglomerado es menor que la distancia entre cualquier punto del conglomerado y otro cualquiera que no pertenezca al mismo.

De todas formas, existen conceptos fundamentales como los de conglomerado o similitud, que,como dijo BONNER (1964), "el último criterio para evaluar el significado de tales términos es el juicio de valor del usuario" y, como afirma CORMACK (23) existen muchas ideas intuitivas de lo que constituye un conglomerado, pero pocas definiciones formales, aunque siempre se incluyen dos ideas básicas: COHESION interna y AISLAMIENTO externo.

Otros autores definen diferentes tipos de conglomerados. Por ejemplo, JOHNSON (72) considera entidades que son óptimamente compactas (en algún sentido) como un tipo de conglomerado y entidades que están conectadas óptimamente como otro tipo. Estos dos tipos de conglomerados parecen similares a los que describieron CATTELL y COULTER (23) como (HOMOSTATS y SEGREGATES) homostáticos y segregados, donde un homostático es un conjunto de individuos no generalmente similares, mientras que un segregado es una serie de individuos con similares por encima de la media.

VAN RYSBERGEN (109) define conglomerado L^s como subconjuntos de entidades en las que toda similitud, dentro del conglomerado, es menor que toda similitud desde dentro del subconjunto hacia fuera de él.

Una descripción de lo que constituye un conglomerado y que se aproxima a nuestra idea intuitiva del término, puede ser el considerar las entidades como puntos de espacio p-dimensional, en donde cada una de las p variables están representadas por los ejes de dicho espacio. Los valores de la variable, para cada entidad, nos definen ahora una coordenada p-dimensional de este espacio. Los conglomerados pue-

den, de esta forma, describirse como regiones continuas de este espacio que contienen una, relativamente, alta densidad de puntos, separadas de otras regiones por regiones que contienen una, relativamente, baja densidad de puntos. Los conglomerados así descritos, se suelen denominar conglomerados naturales.

VII. PROCEDIMIENTOS Y ALGORITMOS: ANALISIS DE CONGLOMERADOS:

El término más utilizado para las técnicas que intentan separar datos constituyendo grupos es el ANALISIS DE CONGLOMERADOS que desempeña un importantísimo papel en la ciencias taxonómica y del que pretendemos hacer, a continuación, una breve exposición.

KENDALL y STUART (42) propusieron que el término análisis de conglomerados se utilizara cuando se intentara agrupar variables y clasificación para agrupar individuos.

El primer trabajo en el que se describe lo que entendemos por ANALISIS DE CONGLOMERADOS fue debido a TRYON (1939).

Con posterioridad, los principales investigadores de las técnicas de conglomeración han sido WILLIAM y LAMBERT (1959), CASETTI (1964), BALL y HALL (1965), EDWARDS y CANALLI-SFORZA (1965), GOWER (1966), JOHNSON (1968), HARRISON (1968), FISHER (1968), GREEN y RAD (1969), FLEISS y ZUBIN (1969), BOLSHEV (1969), COLE (1969), HUNG y RUBES (1970), RAND (1970), JARDINE y SIBSON(1971), BALL (1971) KAND y HOLLEY (1972).

A.- EL PROBLEMA DE LA CONGLOMERACION

Consideremos una población de n individuos que representamos por $I = (I_1 \ldots \ldots I_n)$.

Suponemos que existe un conjunto de caracteres $C = (C_1 \ldots \ldots C_p)$ que hemos observado, y que pueden poseer los anteriores individuos.

Sea x_{ij} la medida del carácter i para el individuo j y sea $X_j = (x_{ij})$ el vector pxl de tales medidas; de esta forma, obtenemos un conjunto de pxl vectores de medida, $X = (X_1, X_2 \ldots X_n)$ que nos describen el conjunto I.

El conjunto X puede ser considerado como n puntos de un espacio p-dimensional Euclideo, Ep.

Entonces, el importantísimo problema de la conglomeración puede plantearse de la siguiente manera: "Si m es un entero menor que n, hemos de determinar m subconjuntos, que llamaremos CONGLOMERADOS (CLUSTERS) de individuos de $I, I_1, I_2 \ldots I_m$, tales que I_i pertenezca a uno y solo uno de dichos subconjuntos, de modo que aquellos individuos que sean asignados al mismo conglomerado, sean similares, mientras que los diferentes conglomerados no sean similares".

Una solución al problema anterior consiste, generalmente, en determinar una partición que satisfaga algún criterio de optimalidad, criterio que puede darse en términos de una relación funcional que refleje los niveles de deseabilidad de las diversas

particiones o agrupaciones. Esta relación funcional se denomina FUNCION OBJETIVO.

Tal y como ha quedado planteado el problema de la conglomeración, se hace necesario definir el término SIMILITUD y, por tanto, el de NO similitud que denominaremos DISIMILITUD, cosa que haremos más adelante.

B.- DIVERSAS DEFINICIONES DE CONGLOMERADO

Según el Diccionario de términos estadísticos de KENDALL y BUCKLAND, un CONGLOMERADO es "un grupo (conjunto) de elementos contiguos de una población estadística". Naturalmente, habría que determinar la idea de contigüidad, palabra clave de la definición anterior.

Otras definiciones del término "conglomerado" son:

- "Un conglomerado es un conjunto de entidades que son semejantes y entidades de diferentes conglomerados no son semejantes".

- "Entidades de un mismo conglomerado son más similares unas a otras que entidades de diferentes conglomerados".

- Según WALLACE y BOULTON (23), "un conglomerado es un subconjunto de entidades que se pueden considerar equivalentes en alguna discusión".

- GENGRELLI (51) define un conglomerado como una agregación de puntos de un espacio "prueba" tal que la distancia entre dos puntos del conglomerado es menor que la distancia entre todo punto del conglomerado y todo punto que no pertenezca a dicho conglomerado.

- Según GEOFFREY H. BALL, en su "Glosario de términos comunes" (11), un conglomerado es un conjunto de objetos similares y añade que "intuitivamente es una colección de objetos que son similares unos a otros y no demasiado similares a otros objetos

del conjunto de datos".

- JARDINE (70) define un conglomerado como un subconjunto del conjunto de objetos tal que los objetos interiores al conglomerado están más cercanos entre sí que los objetos que no están en él.

Se observa claramente en todas las definiciones anteriores que en cada una de ellas figura una palabra clave que, a su vez, necesitaría definirse para clarificar las ideas.

Siguiendo a EVERITT (11)"no existe un acuerdo universal sobre lo que es un conglomerado: y de hecho es, probablemente, cierto que ninguna definición aislada sea suficiente".

BONNER (42) sugirió que "el criterio final para evaluar el significado de términos tales como conglomerado o similitud es el juicio de valor del usuario".

Vemos que existen, como afirma CORMACK, "muchas ideas intuitivas, a menudo conflictivas, de lo que constituye un conglomerado, pero pocas definiciones formales".

Lo que sí es cierto es que en todo lo que se puede afirmar y comentar acerca de la idea de conglomerado, subyacen dos aspectos fundamentales: COHESION INTERNA y AISLAMIENTO EXTERNO. 7*

A veces, se hace hincapié en el aislamiento:

ROGERS (23) encontró la máxima restricción aceptable:

entidades similares no serán incluídas en diferentes conglomerados y, además, se observaría una discontinuidad entre los conglomerados.

A veces, se destaca el término cohesión:

CATTELL (23) decía que: "se incluirá un individuo en un conglomerado si su más baja (pequeña) correlación con todo miembro, es mayor que algún umbral (THRESHOLD).

NEEDHAM (23): "la suma de las similitudes de todo miembro con otro miembro, sería mayor que la suma de sus similitudes a los no miembros y, recíprocamente, para los no miembros".

Otros autores definen varios tipos diferentes de conglomerado. Así, JOHNSON (72) considera entidades que son óptimamente COMPACTAS (en algún sentido) como un tipo de conglomerado y entidades que están óptimamente CONECTADAS como otro tipo.

Estos dos tipos de conglomerado parecen ser similares a los que describieron CATTELL y COULTER (23) y a los que denominaron HOMOSTATIVOS y SEGREGADOS donde un conglomerado homostático es un conjunto de individuos inusitadamente (insólitamente) similares, mientras que un conglomerado segregado es una serie de individuos con altas similaridades medias. Estas ideas formalizan la distinción que hace SNEATH entre conglomerados COMPACTOS y conglomerados DISEMINADAMENTE LARGOS.

EVERITT (42) nos da una descripción muy clara de lo que constituye un conglomerado: se consideran los objetos como puntos

de un espacio p-dimensional en el que cada una de las p-variables se representa en cada uno de los diferentes ejes de este espacio. Los distintos valores para cada objeto, nos definen una coordenada p-dimensional o p-coordenadas en este espacio. Los conglomerados pueden describirse ya como regiones continuas de este espacio que contienen una densidad de puntos relativamente alta separados de otras regiones por zonas que contienen una densidad de puntos relativamente baja. A estos conglomerados, EVERITT los llama NATURALES.

Lo que sí es evidente, y es una idea intuitiva en todos los que trabajan en este campo, es que los miembros de un conglomerado son más próximos o cercanos o afines unos a otros que al resto de los individuos, pero las peculiaridades de esta relación no son descifradas.

Para definir los conglomerados, también, se recurre, a veces, al estudio de sus parámetros como pueden ser, por ejemplo, la densidad de OTUS en el hiperespacio de atributos, "el volumen ocupado" por el conglomerado, la conexión existente entre los miembros de un mismo conglomerado, los espacios entre conglomerados adyacentes comparados con los diámetros de dichos conglomerados, etc.

C.- FUNDAMENTOS TEORICOS DEL ANALISIS DE CONGLOMERADOS

C.1. INTRODUCCION:

Según afirma CORMACK (23), "se han desarrollado muchas técnicas de ANALISIS DE CONGLOMERADOS pero sin base formal alguna".

Por tal razón, vamos a intentar, en este apartado, dar una fundamentación teórica de dicho análisis con vistas, además, al estudio de las técnicas jerárquicas, las más utilizadas, y que nos permitirán, más adelante, entrar en el estudio comparativo de algunas métricas.

Ya hemos dado diferentes definiciones de lo que se entiende por ANALISIS DE CONGLOMERADOS que, resumiendo todas aquellas definiciones, es un conjunto de métodos y técnicas para resolver importantísimos problemas de clasificar elementos, individuos u objetos, problema que, en términos generales, consiste en colocar o adscribir dichos elementos en grupos o clases, de tal modo que los que pertenezcan a una clase sean más cercanos en algún sentido, es decir, conseguir clases naturales o conglomerados.

El hecho frecuente de que no exista una clasificación "a priori" de los elementos, nos lleva al hecho de que el análisis de conglomerados es, fundamentalmente, una herramienta para la exploración de los datos, es decir, estudiamos los datos para ver si, efectivamente, existen agrupaciones útiles y naturales.

Pensemos, a modo de ejemplo, en diversos problemas: el biólogo, que pretende clasificar bacterias a partir de los resulta-

dos de un conjunto de pruebas; el meteorólogo que desea agrupar o definir zonas climáticas a partir de una colección de datos; el botánico que pretende conseguir fenones, basándose en los lazos de parentesco de las especies, y así sucesivamente.

Es bastante sorprendente observar qué problemas tan diferentes como los anteriores posean rasgos y características comunes y, aún lo es más, el hecho de que los datos de cada problema pertenecen a una misma clase general, es decir, un conjunto de objetos y otro de descripciones que están basados o constituídos por una colección de características.

El primer paso de todo problema clasificatorio es definir con precisión y conocer lo mejor posible los objetos o entidades, OTUS, a clasificar. Dichos objetos estarán definidos por sus respectivos caracteres y los podemos disponer en la forma de una matriz, de datos.

Los OTUS se representan en un espacio cuyas dimensiones son los caracteres, un espacio de n dimensiones, por ejemplo, (para n caracteres) pero a causa de las correlaciones de los caracteres de los OTUS puede, generalmente, reducirse a otro espacio de menor número de dimensiones, con pequeña pérdida de información, puesto que de esta forma es más fácil y cómodo el trabajo y, además, eliminamos aquellos caracteres que sean ineficaces. Ahora bien, aunque el objetivo sea reducir la dimensionalidad, ha de mantenerse, tanto como sea posible, la ESTRUCTURA de las observaciones.

Para obtener ese nuevo espacio de menor número de dimensiones necesitamos:

a) una medida de la aproximación de los dos espacios.

b) un algoritmo para encontrar el nuevo subespacio que optimice la medida anterior.

Para resolver el apartado b) se utilizarán las transformaciones lineales que permiten eliminar los caracteres menos importantes; el procedimiento más utilizado es el de las "componentes principales".

Para la búsqueda del nuevo espacio, se suelen utilizar, entre otros, los dos criterios siguientes: (HAND)

1. Se elige el hiperespacio que maximice la suma de cuadrados entre las distancias muestrales:

$$\sum_i \sum_j (\bar{x}_i - \bar{x}_j)'(\bar{x}_i - \bar{x}_j)$$

2. La función de entropía que mide la uniformidad de las proyecciones sobre los ejes del hiperespacio de dimensión d',

$$-\sum_{i=1}^{d} p_i \log p_i$$

siendo p_i la media muestral de los cuadrados de las proyecciones de los x_j sobre la nueva i-ésima coordenada. [8]

Según BARTKO (2): "a menos que exista una pronunciada estructura de conglomeración, los resultados de un análisis de conglomerados aplicados a las variables originales, pueden diferir

marcadamente de un análisis similar aplicado al espacio definido por las nuevas componentes".

Esto es, naturalmente, un riesgo, sin embargo, la reducción de la dimensionalidad no lleva implícito un sacrificio de datos.

También se han aplicado en esta linea los métodos no lineales. Las transformaciones no lineales de la muestra del espacio original, de dimensión d, al nuevo espacio de dimensión d', d'< d, se han basado en diversos criterios estructurales.

Uno de los más frecuentes es el criterio del STRESS, de KRUSKAL (2) definido por

$$\sum_{i<j} (D_{ij}-D'_{ij})^2 \;/\; \sum_{i<j} D_{ij}^2$$

donde D_{ij} es la distancia (o disimilitud)entre los objetos x_i y x_j del espacio d-dimensional y D'_{ij} es la correspondiente distancia del espacio d'-dimensional.

Otro criterio es el de SAMMON (2):

$$\frac{1}{\sum_{i<j} D_{ij}} \sum_{i<j} (D_{ij}-D'_{ij})^2 \;/\; D_{ij}$$

C.2. FUNCIONES DE COMPARACION: SIMILITUDES Y DISIMILITUDES:

"Los métodos basados en la idea de similitud o disimili-

tud son considerados como el grupo más importante de técnicas para analizar matrices de objetos por atributos" SIBSON (109).

La elección de esa medida de similitud o disimilitud entre los OTUS, que ya constituyen la matriz de datos, constituye, junto a la técnica de conglomeración conveniente, el nudo gordiano del análisis de conglomerados.

Todas estas medidas tienen algo en común, y es que están, en alguna forma, derivadas de la información obtenida a partir de los objetos a clasificar, bien sean individuos o clases de individuos.

Los términos SIMILITUD y DISIMILITUD son ambiguos, puesto que pueden cubrir, al menos, los siguientes conceptos distintos JARDINE (70):

a) A-SIMILITUD, que se aplica a individuos o a clases de individuos que no VARIAN con respecto a los atributos considerados, es decir, es el concepto de similitud asociado con la posesión de propiedades comunes. Este tipo de medidas se pueden transformar en medidas de atributos, pero no al contrario.

   Pueden distinguirse dos clases de medidas de similitud, según que dependan o no del control preliminar de los estados de cada atributo.

b) I-DISTINGUIBILIDAD, es la disimilitud asociada con la diagnosis de individuos, o sea, la extensión a qué

clase de individuos pueden ser distinguidos y que puede contemplarse en términos de la probabilidad de reasignación de un individuo de una de las clases sobre una base de información relativa a sus estados de atributos. Una medida de distinguibilidad es, generalmente, un coeficiente de disimilitud.

c) D-DISIMILITUD, concepto muy sutil y que depende de la información suministrada por el hecho de que una clase de individuos se adscriba a una o a otra de las diferentes clases o tipos ya establecidos. La cantidad de información dada de esta forma no está bien definida, puesto que puede diferir de los resultados de la identificación. La disimilitud es un valor típico y adecuado de la información que se espera lograr para tal identificación.

Los conceptos anteriores dan lugar a tres clases de medidas que, globalmente, llamaremos FUNCIONES DE COMPARACION y que son funciones de $X \times X \longrightarrow \mathbb{R}$ siendo X el conjunto de OTUS a clasificar y $\mathbb{R}$ el conjunto de los números reales.

Tales funciones se representan mediante matrices nxn cuyas filas y columnas corresponden a los n elementos del conjunto X a clasificar.

Es evidente que,

$f(x,x) \geq f(x,y)$ para SIMILITUDES

$f(x,x) \leq f(x,y)$ para DISIMILITUDES

siendo x,y dos elementos cualesquiera de X.

Si la función de comparación es simétrica entonces será: $f(x,y) = f(x,y)\ \forall x,y \in X$, y si f mide SIMILITUDES, -f medirá DISIMILITUDES y recíprocamente:

DEFINICION: Una función de comparación con valores reales no negativos, S, se dice que es una medida de SIMILITUD si se verifican las siguientes condiciones:

1. $0 \leq S\ (x,y) < 1$ para $x \neq y$
2. $S(x,y) = 1$ si y solo si $x = y$
3. $S(x,y) = S(y,x)$

A partir de las similitudes de los n elementos del conjunto X, se construye la matriz nxn, o matriz de similitud, que es simétrica y cuyos elementos son coeficientes de similitud.

DEFINICION: Un <u>coeficiente de disimilitud</u> es una función de comparación $d: X \times X \longrightarrow R^{+} \cup \{0\}$ tal que

1. $d(x,y) \geq 0 \quad \forall x,y \in X$
2. $d(x,x) = 0 \quad \forall x \in X$
3. $d(x,y) = d(y,x) \quad \forall x,y \in X$

A veces, los coeficientes de disimilitud pueden tener alguna o algunas de las propiedades siguientes:

4. UNIFORMIDAD:
   $d(x,y) = 0 \Rightarrow d(x,z) = d(y,z) \quad \forall\, x,y,z \in X$
   y entonces, los coeficientes de disimilitud se llaman UNIFORMES.

5. PRECISION:
   $d(x,y) = 0 \Rightarrow x=y \;,\; \forall\, x,y \in X$
   Se dice, entonces, que d es PRECISO.

6. DESIGUALDAD TRIANGULAR (METRICA):
   $d(x,y) + d(y,z) \geq d(x,z), \quad \forall\, x,y,z \in X$
   a tales coeficientes se les llama METRICOS o METRICAS.

7. DESIGUALDAD ULTRAMETRICA:
   $d(x,z) \leq \max \{d(x,y), d(y,z)\} \quad \forall\, x,y,z \in X$
   cuando d, cumple la condición anterior, se dice que es ULTRAMETRICO.

Es importante observar que un coeficiente de disimilitud no tiene por qué satisfacer la condición de PRECISION puesto que tal cosa podría ser un reflejo del hecho de que los objetos distintos pudieran coincidir en sus descripciones.

Hay otro hecho mucho más significativo para un coeficiente de disimilitud, como es el que no necesita satisfacer la desigualdad métrica: "estamos acostumbrados, SIBSON (109), a pensar en las disimilitudes como algo próximo a las distancias y, por ello, es sorprendente que la desigualdad métrica, tan importante en la geometría de las distancias, sea omitida en la definición de un coeficiente de disimilitud".

La razón de la omisión es que la desigualdad métrica nos fuerza a considerar la adición de valores de un coeficiente de disimilitud, cosa que, no nos está permitido hacer. Por tanto, las disimilitudes pueden considerarse como el polo opuesto del espectro de las distancias geométricas.

Según ESCUDERO (41), las cualidades que ha de tener una apropiada medida de disimilitud entre dos elementos x e y con f características son

a'. La disimilitud será positiva para los elementos distintos: $d(x,y/\ 1,2,\ \ldots,\ f) > 0$

b'. La disimilitud de un elemento consigo mismo es nula: $d(x,x/\ 1,2,\ \ldots,\ f) = 0$

c'. La disimilitud entre x e y no debe estar afectada por la denominación de los elementos: $d(x,y/\ 1,2,\ \ldots,f) = d(x,y/\ 1,2,\ \ldots,f)$

d'. La disimilitud debe ser aditiva para características independientes: $d(x,y/\ 1,2,\ \ldots,f) = \sum_{i=1}^{f} d\ (x,y/\ f)$

e'. La disimilitud no debe decrecer al añadir más características: $d(x,y/\ 1,2\ \ldots\ f) \nleq d(x,y/\ 1,2,\ \ldots,f,\ f+1)$

f'. La disimilitud debe ser invariante a rotaciones y transformaciones.

g'. La disimilitud debe tener en cuenta la interdependencia de las características en el grupo en el que están los elementos.

h'. La disimilitud debe ser sensible a la diferente ponderación a introducir en la cuantificación de la divergencia de cada característica, de acuerdo con su mayor o menor importancia en la discriminación de los elementos

D.- TECNICAS DE CONGLOMERACION

Han sido desarrolladas muchas técnicas de conglomeración sin base formal, como algoritmos.

En general, y siguiendo a CORMACK (23) se utilizan para la conglomeración tres tipos de procedimientos:

a. AGLOMERATIVOS: Una serie de fusiones sucesivas de los n objetos en grupos.

b. DIVISIVOS: Partición del conjunto completo E, sucesivamente en otras particiones más finas.

c. CONGLOMERACION: Sucesivas recolocaciones de individuos entre los conjuntos de alguna partición inicial.

Los métodos a. y b. representan los datos como un DENDROGRAMA en el que los conglomerados se obtienen cortando a distintos niveles. Los procedimientos del apartado c. se utilizan para hallar directamente una partición de E con propiedades que se aproximan a algún deseo o motivo.

Según EVERITT (42) las técnicas o procedimientos de análisis de conglomerados son los siguientes:

a. TECNICAS JERARQUICAS en las que las clases, a su vez, son clasificadas en grupos repitiéndose el proceso a diferentes niveles para formar un árbol.

b. TECNICAS DE OPTIMACION-PARTICION según las cuales, las clases se forman por optimación de un criterio de conglomeración. Las clases son mutuamente excluyentes, formando así una partición del conjunto de entidades.

c. TECNICAS DE DENSIDAD O INVESTIGACION MODAL en las que los conglomerados se forman buscando regiones que contengan una concentración relativamente densa de entidades.

d. TECNICAS DE FORMACION DE GRUPOS cuyas clases o bloques puedan superponerse.

e. OTRAS, que son métodos que no encajan claramente en alguna de las anteriores

Los mencionados grupos no se excluyen unos a otros y se pueden colocar varias técnicas de conglomeración en más de una categoría.

Los OBJETIVOS de los usuarios de las técnicas de conglomeración son, frecuentemente, distintos.

BALL (12) nos indica siete usos posibles de dichas técnicas:

a. Encontrar una verdadera tipología

b. Modelo adecuado

c. Predicción basada en grupos

d. Prueba de hipótesis

e. Exploración de datos

f. Generación de hipótesis

g. Reducción de los datos

Como nos dice AZORIN POCH, hemos de buscar una CLASIFICACION que sea,

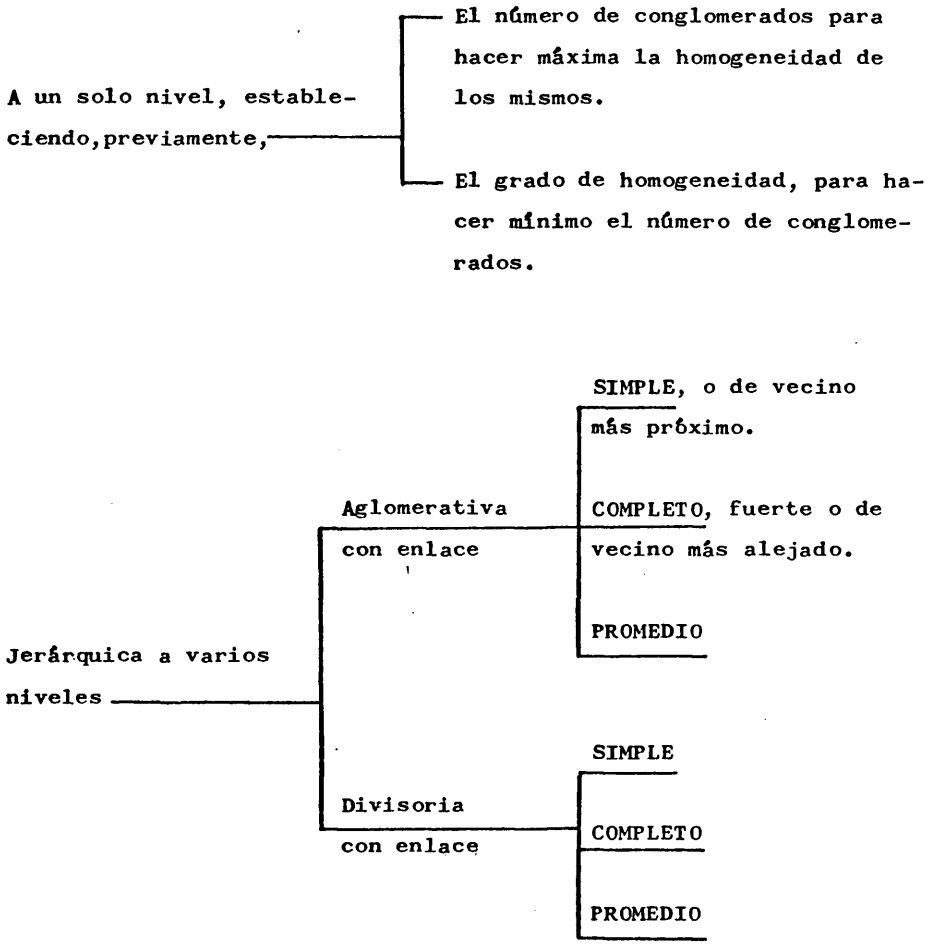
- La más HOMOGENEA, constituída por conglomerados de elementos más parecidos entre sí, entre las que son igualmente económicas, en el sentido de tener el mismo número de conglomerados o clases.

- La más ECONOMICA, es decir, la que tenga menor número de clases entre las que tienen un mismo grado de homogeneidad.

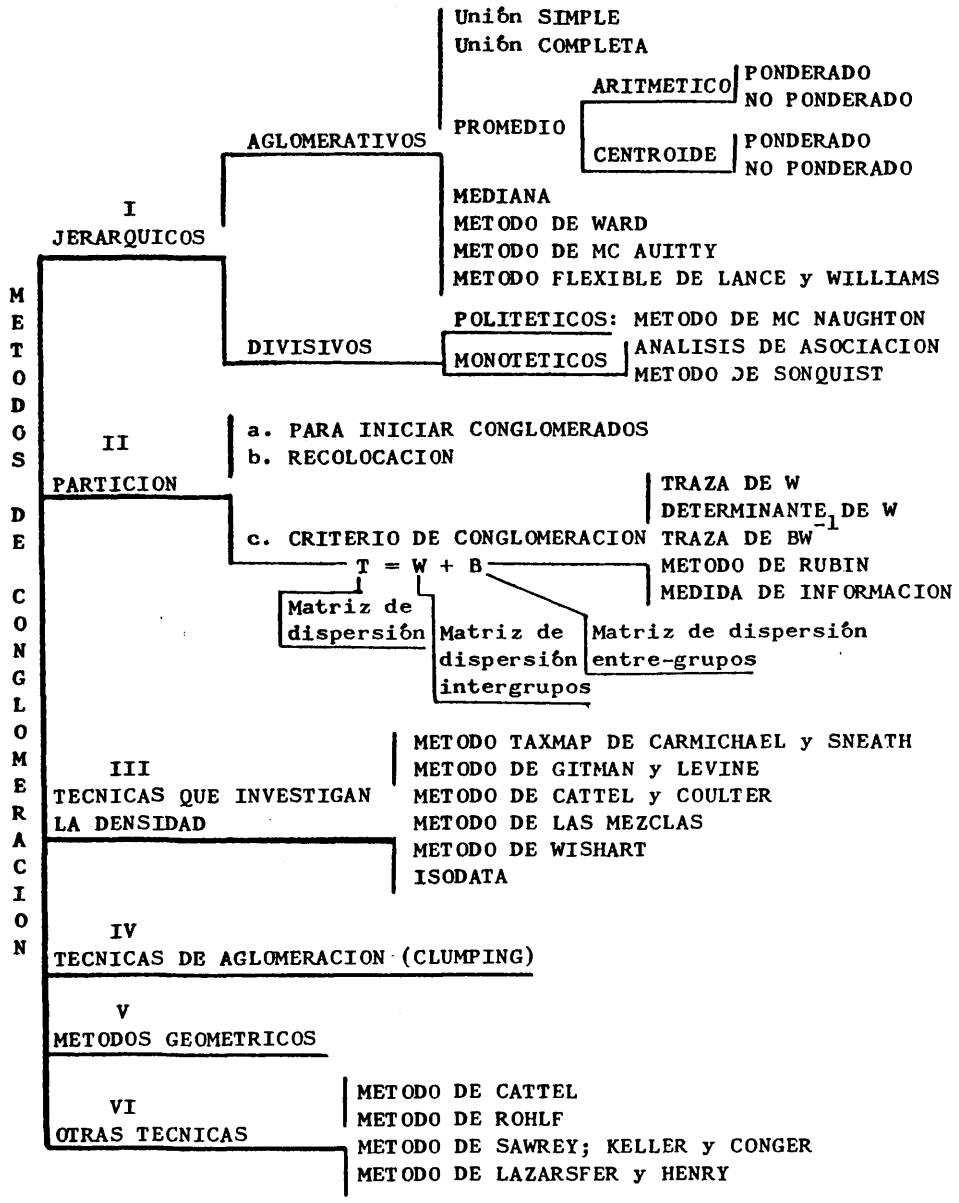
- La OPTIMA, o al menos aceptable para el par dual: homogeneidad y economía.

Un esquema general de los procedimientos de clasificación o conglomeración es el siguiente:

CONGLOMERACION



Como resúmen de los diversos métodos que hemos estudiado a lo largo de la bibliografía relacionada con este tema, ofrecemos un esquema, a continuación, de los diversos métodos de conglomeración.


METODOS DE CONGLOMERACION
I
JERARQUICOS
AGLOMERATIVOS
Unión SIMPLE
Unión COMPLETA
PROMEDIO
ARITMETICO
PONDERADO
NO PONDERADO
CENTROIDE
PONDERADO
NO PONDERADO
MEDIANA
METODO DE WARD
METODO DE MC AUITTY
METODO FLEXIBLE DE LANCE y WILLIAMS
DIVISIVOS
POLITETICOS: METODO DE MC NAUGHTON
MONOTETICOS
ANALISIS DE ASOCIACION
METODO DE SONQUIST
II
PARTICION
a. PARA INICIAR CONGLOMERADOS
b. RECOLOCACION
c. CRITERIO DE CONGLOMERACION
TRAZA DE W
DETERMINANTE DE W
TRAZA DE BW-1
T = W + B
METODO DE RUBIN
MEDIDA DE INFORMACION
Matriz de dispersión
Matriz de dispersión intergrupos
Matriz de dispersión entre-grupos
III
TECNICAS QUE INVESTIGAN LA DENSIDAD
METODO TAXMAP DE CARMICHAEL y SNEATH
METODO DE GITMAN y LEVINE
METODO DE CATTEL y COULTER
METODO DE LAS MEZCLAS
METODO DE WISHART
ISODATA
IV
TECNICAS DE AGLOMERACION (CLUMPING)
V
METODOS GEOMETRICOS
VI
OTRAS TECNICAS
METODO DE CATTEL
METODO DE ROHLF
METODO DE SAWREY; KELLER y CONGER
METODO DE LAZARSFER y HENRY

D.1. CARACTERISTICAS FUNDAMENTALES DE LOS GRANDES GRUPOS DE METODOS DE CONGLOMERACION:

1. METODOS JERARQUICOS

Tienen por objeto formar conglomerados finales, agrupando subconglomerados iniciales, es decir, se puede empezar con n subconglomerados de un punto cada uno y combinarlos para formar otros mayores que, a su vez, se vuelven a combinar para conseguir otros nuevos y así sucesivamente hasta conseguir un número determinado de conglomerados. Naturalmente, en cada paso se van combinando aquellos grupos que sean más similares.

El proceso anterior se puede efectuar al revés, es decir, comenzando con un conglomerado que contiene todos los puntos y dividirlo en dos que, a su vez, se vuelven a dividir, y así sucesivamente.

Los dos procedimientos anteriores, que son jerárquicos y complementarios, se denominan respectivamente, AGLOMERATIVOS y DIVISIVOS.

La representación más sencilla de los resultados de un análisis jerárquico de conglomeración es el DENDROGRAMA, cuyo eje vertical proporciona una escala para medir la distancia o disimilitud entre dos conglomerados.

2. METODOS DE PARTICION

El objeto de estos métodos es la partición del conjun-

to de n objetos en un determinado número de conglomerados disjuntos de tal modo que cada objeto o individuo pertenezca a un grupo.

3. METODOS DE DENSIDAD

Cada objeto de alta densidad se utiliza como centro de un conglomerado al que se van añadiendo aquellos puntos cercanos de alta densidad.

4. METODOS DE AGLOMERACION (CLUMPING)

Así como en los métodos de partición, los conglomerados no tienen elementos comunes, en éstos métodos los conglomerados se solapan de tal modo que cada objeto pertenece, al menos, a un

A tal división del conjunto en aglomerados, se le llama RECUBRIMIENTO del conjunto de datos.

5. METODOS GEOMETRICOS

En estos métodos cada objeto se representa por un punto de un espacio, que suele ser de dos o tres dimensiones. Esta representación geométrica tiene la propiedad de que aquellos elementos que sean más similares, se representan por puntos más cercanos.

D.2. CUALIDADES DE LOS METODOS DE CONGLOMERACION:

A continuación, exponemos ocho cualidades que, teoricamente, pueden tener un método de conglomeración y que, el hecho de

poseer unas y no otras, marcan las diferencias substanciales entre unos y otro métodos, (SNEATH (110) y ESCUDERO (40).

Pueden dar lugar a 2^8 tipos diferentes de métodos de los que algunos son probablemente imposibles.

1a. METODOS AGLOMERATIVOS

Se parte de n grupos y se pretende llegar al grupo total procediendo, en cada nivel, a fusionar aquellos dos grupos que sean más similares.

1b. METODOS DIVISIVOS

Partiendo de un grupo formado, al nivel cero, por todos los elementos, se reparten dichos elementos al nivel uno para formar dos grupos, maximizando alguna medida de divergencia preestablecida.

2a. METODOS JERARQUICOS

Tienen por objeto agrupar conglomerados para formar uno nuevo, o bien separar conglomerados formándolos nuevos.

2b. METODOS NO JERARQUICOS

En ellos se escoge una partición inicial de los elementos, y, a continuación se altera dicha partición moviendo los elementos de un grupo a otro buscando una partición mejor. El número inicial de particiones está determinado previamente o su determinación forma parte del método.

3a. METODOS SOLAPADOS

Un individuo puede pertenecer simultaneamente a más de un grupo.

3b. METODOS NO SOLAPADOS (O EXCLUYENTES)

Un elemento solo puede pertenecer a un grupo.

4a. METODOS SECUENCIALES

A cada elemento se le aplica la misma sucesión de operaciones para conseguir los grupos.

4b. METODOS SIMULTANEOS

En estos métodos, con una simple operación se consigue la agrupación en conglomerados de toda la colección de elementos.

5a. METODOS DIRECTOS

Utilizan algoritmos tales que una vez que un elemento ha sido asignado a un grupo, no vuelve a salir del mismo.

5b. METODOS ITERATIVOS

Corrigen sus propias asignaciones volviendo a comprobar en sucesivas iteraciones si la asignación es óptima y de lo contrario se vuelve a efectuar el reagrupamiento.

6a. METODOS ADAPTATIVOS

Estos métodos "aprenden" en su ejecución y van cambiándo de medida o de criterio a optimizar a lo largo del proceso.

6b. METODOS NO ADAPTATIVOS

El método es fijo a lo largo de todo el proceso y el algoritmo se encamina directa o iterativamente hacia la solución.

7a. METODOS PONDERADOS

Los diferentes algoritmos utilizados en estos métodos consideran como más importantes ciertos tipos de relaciones o ciertas dimensiones.

7b. METODOS NO PONDERADOS

Mejor sería llamar a estos métodos, métodos de igual ponderación. En su utilización no se prefiere ninguna dirección, ni ningún elemento. Todas las relaciones tienen la misma importancia.

8a. METODOS QUE UTILIZAN CRITERIOS LOCALES

El procedimiento de conglomeración se lleva a cabo observando en cada nivel la bondad de los resultados e incluso se pueden utilizar distintas métricas en las diferentes partes de la clasificación taxonómica.

8b. METODOS QUE UTILIZAN CRITERIOS GLOBALES

No se tienen en cuenta los resultados a diferentes niveles. Lo importante es el resultado final sin alterar el criterio a lo largo de todo el proceso de conglomeración.

E.- METODOS JERARQUICOS

Un método jerárquico de conglomeración obtiene las clases, grupos o conglomerados finales, objeto del estudio, agrupando jerárquicamente subconglomerados o dividiendo en otros más pequeños, los conglomerados iniciales.

Así, podemos comenzar con n subconglomerados (n es el número de elementos) que contienen un elemento cada uno y los combinamos para formar subconglomerados mayores y después volver a combinar éstos para conseguir otros nuevos y así sucesivamente, hasta obtener el número deseado o hasta que todos los elementos forman un único conglomerado.

En este proceso, en el paso o nivel h, habrá un número c(h) de conglomerados y algunos de éstos se agruparán para formar uno nuevo, de acuerdo con un cierto criterio; en el paso h+1 tendremos c(h+1) conglomerados, verificándose que $c(h+1) \leq c(h)$.

Alternativamente, podríamos comenzar con un único conglomerado que contiene todos los elementos a clasificar y dividirlo para dar lugar a dos nuevos conglomerados más pequeños.

Volviendo a dividir éstos últimos, obtenemos otros más pequeños y seguiremos así sucesivamente hasta conseguir el número deseado, o llegar a tantos conglomerados como elementos.

Los dos procesos anteriores, ambos jerárquicos, son, evidentemente, complementarios y son llamados métodos AGLOMERATIVOS y DIVISIVOS, respectivamente.

Todo el proceso de conglomeración depende y parte de una triz, nxn, de similitudes o de disimilitudes; por tanto, podría-os decir que el INPUT de un proceso jerárquico de conglomeración onsiste, únicamente, en un conjunto de $\frac{n(n-1)}{2}$ medidas de simili-ud o de disimilitud entre los n elementos a clasificar; la elec-ión de esa medida de acercamiento o de separación entre los ele-entos y su generalización a las medidas correspondientes entre onglomerados, constituyen el punto neurálgico del proceso de con-lomeración.

Vamos a exponer, a continuación, un estudio teórico de os métodos jerárquicos numéricamente estratificados, que son los ás frecuentemente utilizados. Estudio que hemos elaborado siguien-o las ideas de JOHNSON (72), JARDINE y SIBSON (70) y GOWER y OSS (35), y que nos llevará al centro de este trabajo de compa-ación de algunas métricas.

El punto de partida de todo método de conglomeración es coeficiente de disimilitud o de similitud.

A lo largo de la siguiente exposición, nos referiremos los primeros, ya que es fácil obtener,a partir de ellos, las imilitudes.

DEFINICION: Un coeficiente de disimilitud sobre un conjunto X de elementos u OTUS es una función: $d: X \times X \longrightarrow \mathbb{R}$ que satisface las siguientes condiciones:

1. $d(x,y) \geq 0 \qquad \forall\, x,y \in X$
2. $d(x,x) = 0 \qquad \forall\, x \in X$
3. $d(x,y) = d(y,x) \qquad \forall (x,y) \in X$

El punto final de un método jerárquico de conglomeración es un DENDROGRAMA o diagrama de árbol.

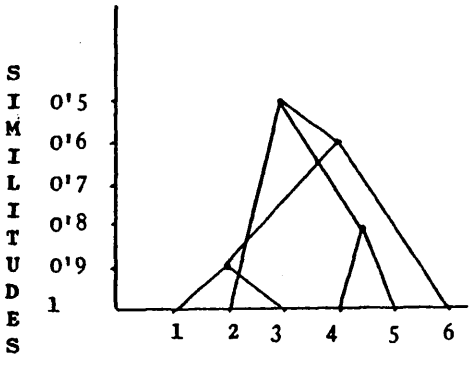
Un dendrograma es una representación diagramática de los resultados de un proceso jerárquico de conglomeración, proceso que es realizado a partir de una matriz de disimilitudes o similitudes.

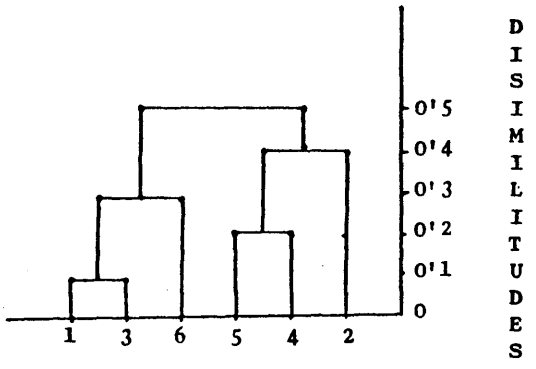
El dendrograma es, por tanto, una estructura que representa gráfica o geométricamente los conglomerados jeráquicos que resultan cuando un proceso de conglomeración opera sobre una matriz de disimilitudes o de disimilitudes.

Hay varias formas de dibujar un diagrama de árbol correspondiente a determinado dendrograma.

En un diagrama de árbol, los objetos pueden colocarse sobre una línea horizontal, y los resultados de la conglomeración, hacia arriba. Los niveles de similitud o de disimilitud a los que se forman los conglomerados, se indican verticalmente.

En la figura siguiente, exponemos el mismo dendrograma pero de diferentes formas, pero teniendo en cuenta que, en todos los diagramas, el sistema de conglomerados a cada nivel debe ser el mismo.


SIMILITUDES
0'5
0'6
0'7
0'8
0'9
1
1 2 3 4 5 6


0'5
0'4
0'3
0'2
0'1
0
1 3 6 5 4 2
DISIMILITUDES

Dada una de las ramas finales del árbol representa los objetos aislados. Cada punto (o segmento) en el que se unen lás ramas tiene asociado un valor numérico, de similitud o disimilitud, d_{ij}, y que es nivel en el que se unen las ramas i y j, el más bajo en el que los elementos i y j pertenecen al mismo conglomerado.

Si seccionamos horizontalmente el diagrama por un nivel cualquiera, obtenemos una partición del conjunto de los elementos iniciales.

Por ejemplo, la linea de puntos trazada en la figura anterior al nivel 0,2 4 nos indica los grupos (1,3), 6,(5,4),2.

Dados n elementos existen muchos árboles que puedan ser construidos a partir del procedimiento jerárquico,pero dada una matriz de disimilitudes o de similitudes, le corresponde únicamente un diagrama de árbol.

Los conglomerados correspondientes a un determinado nivel en un dendrograma, tienen la propiedad de que son disjuntos y todo elemento del conjunto X pertenece a un conglomerado, que puede estar constituído por él solo.

Es decir, los conglomerados forman una PARTICION del conjunto X.

Existe una correspondencia 1-1 entre las particiones de un conjunto y las relaciones de equivalencia definidas sobre él.

Las relaciones definidas en el conjunto X son subconjuntos de X x X.

"Si R es una relación de equivalencia definida en X, entonces los conjuntos $Ca = \{x/(x,a) \in R\}$ forman una partición de X". recíprocamente, "si los conjuntos $\{C_\alpha\}$ forman una partición de X, entonces la relación $R = \bigcup_\alpha C_\alpha \times C_\alpha$ es una relación de equivalencia".

"Las transformaciones de relaciones de equivalencia a particiones y de particiones a relaciones de equivalencia, son inversas".

Al formalizar la idea de dendrograma podemos hacer uso de una relación de equivalencia para describir los conglomerados a cada nivel; los conglomerados son, así las clases de equivalencia de la relación asociada.

¿Qué sucede a diferentes niveles?

Si $h \leq h'$, y C es un conglomerado al nivel h, entonces C está completamente contenido en un conglomerado al nivel h'; se dice entonces que las particiones son ANIDADAS, o ENCAJADAS.

Es evidente que a un determinado nivel, el que fuere, pero que existe, todo elemento del conjunto X pertenece a un solo conglomerado, el conjunto P.

¿Qué ocurre en los niveles donde cambia el sistema de conglomerados?

Vamos a exigir, una vez que tenemos los resultados anteriores, la siguiente condicion:

"el sistema de conglomerados a cada nivel, ha de ser el mismo que a un nivel ligeramente más alto".

Estamos ya en condiciones de dar la siguiente definición formal de DENDROGRAMA:

DEFINICION: Sea E(X) el conjunto de las relaciones de equivalencia definidas en el conjunto X. Un DENDROGRAMA es una función $c:[0,\infty) \longrightarrow E(X)$ que satisface las siguientes condiciones:

1. $0 \leq h \leq h' \Rightarrow c(h) \subset c(h')$
2. $c(h) = X \times X$ para un cierto h.
3. Dado h, $\exists \delta$ tal que $c(h+\delta) = c(h)$

Las anteriores condiciones que ha de cumplir un dendrograma se pueden escribir, más claramente, así:

1. Todo conglomerado a un nivel dada h' es una unión de conglomerados del nivel h, siendo $0 \leq h \leq h'$.

2. Para un h, suficientemente grande, $0 \leq h \leq h'$.

3. Dado h, existe un $\delta > 0$ tal que las conglomeraciones a los niveles h y $h+\delta$ son exactamente iguales.

Hasta ahora no hemos requerido que todos los elementos de X sean distintos al nivel cero. Los dendrogramas con esta

ropiedad se llamar PRECISOS, es decir,

"Diremos que un dendrograma es PRECISO cuando satisface la siguiente condición:
$c(0) = \Delta X$, siendo $\Delta X = \{(x,x)(x \in X)\}$ ".

Ya podemos contemplar los métodos de conglomeración como ransformaciones de coeficientes de disimilitud en dendrogramas.

Ahora bien ¿es posible que esas transformaciones sean unciones?

De esta forma podríamos asociar a cada coeficiente de isimilitud un dendrograma y conseguir:

a. Dado un determinado conjunto de elementos y con un determinado coeficiente de disimilitud, siempre se obtendrá el mismo dendrograma.

b. Todo método de conglomeración numéricamente estratificado, será representable por una función que aplica el conjunto de coeficientes de disimilitud definidos en X en el conjunto de dendrogramas de X.

Este problema se resuelve de la siguiente forma:

1. Si c es un dendrograma, define una aplicación H_c asociada al dendrograma C así:

$$(H_c)(x,y) = \inf\ (h/(s,y) \in c(h))$$

aplicación que es una correspondencia 1 - 1 entre el

conjunto de dendrogramas de X y algún subconjunto del conjunto de los coeficientes de disimilitud de X.

2. Define una relación, T, simétrica y reflexiva, tal que si d es un coeficiente arbitrario de disimilitud,

$$(Td)\ (h) = \{(x,y)/d(x,y) \leq h\}$$

3. Demuestra que Td es un dendrograma si y, solo si (Td)(h) es una relación de equivalencia para todo $h \geq 0$.

4. Y finalmente, llega a que "T es una correspondencia 1-1 entre el conjunto de las ULTRAMETRICAS de X" y "H" es su inversa", por tanto,

" Como T y H definen una correspondencia natural 1-1 es posible IDENTIFICAR el conjunto de los dendrogramas de X con el conjunto de las ultramétricas de X"; de esta forma "podemos considerar un método numéricamente estratificado como una función del conjunto de los coeficientes de disimilitud del conjunto X en el conjunto de las ultramétricas de dicho conjunto".

Es decir, "un método de conglomeración es un proceso que transforma un coeficiente de disimilitud en una clase especial, las ultramétricas, de coeficientes de disimilitud".

JOHNSON (72) considera esquemas jerárquicos de conglomeración y establece una correspondencia entre tales esquemas y las ultramétricas.

Consideremos la sucesión de conglomeraciones $C_o, C_1, \ldots C_m$ y a cada una de ellas le asociamos un número h_j, $j=0,1, \ldots m$.

Existe una conglomeración (conjunto de conglomerados) C_j asociada con cada nivel h_j.

La conglomeración C_o contiene n (hay n elementos) conglomerados de un único elemento cada uno, al nivel cero, $\alpha_o = 0$.

Luego, se verifica que $h_o \leq h_1 \leq h_2 \leq \ldots \leq h_m$ y cada conglomerado de la conglomeración C_j es la unión de conglomerados de la conglomeración C_{j-1}, es decir, si C es un conglomerado al nivel h, C está completamente contenido en un conglomerado C' al nivel h'.

El esquema así definido se denomina ESQUEMA DE CONGLOMERACION JERARQUICA.

Sea el siguiente esquema:

$$\begin{array}{ccccccc} C_o, & C_1, & \ldots & \ldots & , & C_m \\ \downarrow & \downarrow & & & & \downarrow \\ h_o, & h_1, & \ldots & \ldots & , & h_m \end{array}$$

Definimos la siguiente métrica entre los elementos de X:

$d(x,y) = \alpha_i$, siendo i el menor entero del conjunto $(0,1,2, \ldots \ldots, m)$ tal que $x \; C_i \; e \; C_i$.

PROPOSICION.- "d no sólo es una métrica, sino que es una ultramétrica, ya que verifica la propiedad

$$d(x,z) \leq \max \left\{ d(x,y) \ , \ d(y,z) \right\}$$ "

Ver JOHNSON (72)

Sea D(d) la matriz nxn cuyos elementos son las distancias d(x,y) entre todos los pares de elementos del conjunto X, entonces,

"Dado un esquema jerárquico de conglomeración, le corresponde una métrica d".

y recíprocamente,

"Dada la matriz D(d) se puede reconstruir el correspondiente diagrama de árbol".

Es evidente que aquellas tres condiciones que JARDINE imponía al dendrograma en su definición, son similares a las que impone JOHNSON.

La única diferencia entre ambos procedimientos está en que la definición de JARDINE y SIBSON no exige que los objetos sean distintos al nivel h = 0.

GOWER y ROSS (35) discuten la idea del MINIMO ARBOL EXTENDIDO:

DEFINICION. Dados n puntos, un árbol extendido sobre sobre estos puntos es cualquier conjunto de segmentos que unen pares de puntos tales que,

a. no existen recintos cerrados
b. por cada punto pasa,al menos, un segmento
c. el árbol está totalmente conectado.

Las ideas de esta definición proceden de la teoría de grafos.

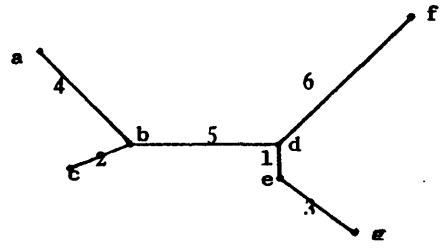
"La longitud de un árbol es la suma de las longitudes de los segmentos que constituyen el árbol".

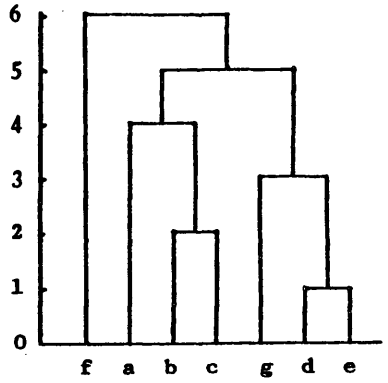
"El MINIMO árbol extendido es el árbol de menor longitud".

Sean $(d_1, d_2, \ldots \ldots, d_n)$ las longitudes de los segmentos de un mínimo árbol, donde h es el número de segmentos.

Un dendrograma puede derivarse del conjunto de los d_i agrupando aquellos dos puntos que están unidos por el menor segmento y procediendo como en una conglomeración de unión simple.

En la siguiente figura, podemos observar el mínimo árbol extendido correspondiente a los mismos elementos para los que se ha dibujado, también, el dendrograma,


a
4
b
5
6
f
c
2
1
d
e
3
g


6
5
4
3
2
1
0
f
a
b
c
g
d
e

Una vez definido lo que es un método de conglomeración, el siguiente paso está relacionado con los métodos de conglomeración numéricamente estratificados.

Para ello se definen, en primer lugar, los CONJUNTOS UNIDOS MAXIMALMENTE:

DEFINICION: Sea R una relación reflesiva y simétrica del conjunto de elementos X, entonces,"un CONJUNTO DE UNION MAXIMAL es un subconjunto $M \subset X$ que satisface las dos condiciones siguientes:

1. $M \times M \; R$
2. $x \notin M \Rightarrow \exists\, y \in M$ tal que $(x,y) \notin R$"

De esta forma un conjunto de unión maxima. es un subconjunto de X en el que la relación R es universal.

Sea, ahora, $\Sigma(X)$ el conjunto de las relaciones reflexivas y simétricas del conjunto X.

Si el procedimiento de conglomeración es jerárquico, el sistema de conglomerados a cada nivel, serán conjuntos de unión maximal en los que la relación R es de equivalencia.

Ya se puede dar la definición de CONGLOMERACION NUMERICAMENTE ESTRATIFICADA, en base a la generalización del dendrograma:

DEFINICION.- Un método de conglomeración numéricamente estratificado es una FUNCION, $C:\{(0,\infty) \longrightarrow \Sigma(X)$ que satisface las siguientes condiciones:

1. $0 \leq h \leq h' < \infty = C(h) \subset C(h')$
2. $C(h)$ es $X \times X$ para un cierto h
3. Dado $h \geq 0$, $\exists \delta$ tal que $C(h+\delta) = C(h)$

Las condiciones son las mismas que las que se impusieron en la definición de dendrograma; la diferencia está en que en aquella, el conjunto final de la aplicación era el conjunto E(X), de las relaciones de equivalencia, definidas en X y en esta definición el conjunto final de la aplicación es el conjunto $\Sigma(X)$ de las relaciones reflexivas y simétricas del conjunto X.

Anteriormente se estableció una correspondencia 1-1 entre los dendrogramas de X y las ultramétricas de X y ahora, aquella correspondencia se extiende a una correspondencia 1-1 entre el conjunto de los métodos numéricamente estratificados de X y el conjunto de los coeficientes de disimilitud sobre X.

Por tanto, se pueden identificar los métodos numéricamente estratificados sobre X con los coeficientes de disimilitud de X y, de esta forma, podemos considerar el método de conglomeración como una función D, que aplica el conjunto de los coeficientes de disimilitud sobre X en algún subconjunto del mismo.

F.- COMENTARIO SUCINTO DE LAS TECNICAS DE CONGLOMERACION MAS UTILIZADAS

Las técnicas más conocidas y más utilizadas son las que se llaman ENLACE SENCILLO y ENLACE COMPLETO o de la DISTANCIA MAXIMA.

F.1. ENLACE COMPLETO O DE LA DISTANCIA MAXIMA:

Este método consiste en lo siguiente:

a) Se construye una matriz simétrica de medidas de distancia para determinar la menor distancia entre pares de objetos.

b) Una vez determinada tal distancia, se agrupan esos individuos en un conglomerado.

c) Se construye una nueva matriz de medidas de distancia seleccionando de entre todas las medidas de esos dos individuos, uno a uno, a todos los otros individuos de la matriz, la mayor medida de distancia. Esta distancia máxima se convierte, entonces, en la medida que representa la relación del conglomerado con los otros objetos o individuos.

d) Se repite el proceso hasta que todos los objetos o individuos y los conglomerados se reunen en un único conglomerado.

Este método puede ser utilizado con similitudes y con medidas de distancia.

Esta estrategia produce conglomerados compactos sin encadenamiento.

F.2. ENLACE SENCILLO O DE LA MINIMA DISTANCIA:

Este método utiliza las mismas operaciones que el anterior, con la excepción de que la distancia seleccionada es la MINIMA y representa la relación de proximidad del nuevo conglomerado,-individuo a cualquier otro conglomerado-individuo.

Este es el más sencillo procedimiento aglomerativo, que puede ser utilizado con medidas de similitud y con medidas de distancia.

La ventaja del enlace SIMPLE es que las sucesivas fusiones de conglomeratos tienen lugar siempre a bajos niveles de similitud entre conglomerados.

Al utilizar este procedimiento, los conglomerados formados están definidos por la condición de que dos entidades E_i y E_j pertenecen al mismo conglomerado si existe una cadena de entidades $E_k, E_l, \ldots \ldots E_q, E_r$ tal que $S_{ik}, S_{ke}, \ldots \ldots, S_{qr}, S_{rj}$ son todos mayores que algún valor inicial t.

WILLIAMS (116) define un "coeficiente de encadenamiento" que expresa la tendencia que las entidades tienen a ser incorporadas a los conglomerados existentes mejor que a construir un nuevo conglomerado.

Los resultados de este método no son satisfactorios si existen entre los conglomerados lo que HODSON denominó "ruidos aleatorios, habiendo sido propuestos diferentes métodos para eliminar dichas perturbaciones. -

F.3.- METODO DEL CENTROIDE:

Fue propuesto por SOKAL y MICHENER (1958) y KING (1966).

Todo conglomerado se considera como un solo punto, su centro de gravedad, CENTROIDE, del espacio euclideo. La distancia entre grupos se define como la distancia entre sus centroides; el procedimiento a seguir consiste en fusionar grupos, de acuerdo con las distancias entre sus centroides, empezando por aquellos que tienen la menor distancia.

Aunque la base geométrica del método sugiere el empleo de la distancia euclidea, también, pueden utilizarse otras medidas de distancia o de similitud.

F.4. METODO DE LA MEDIANA:

Una desventaja del método del CENTROIDE es que si los tamaños de los grupos que se van a fusionar son muy diferentes, el centroide el nuevo grupo estará muy próximo al centroide del grupo mayor e incluso, puede "caer" dentro de él; ello implica que las propiedades características del grupo más pequeño, se perderán virtualmente.

La estrategia del presente método conlleva la independencia de los tamaños de los conglomerados a fusionar suponiendo que todos tienen el mismo y, por tanto, el nuevo grupo estará siempre entre los grupos a fusionar. Además, si representamos los centroides de los grupos a fusionar por (i) y (j), entonces la distancia del centroide de un tercer grupo (h) al grupo formado por la fusión

de (i) y (j) pasa por la mediana del triángulo definido por (i)(j) y (h) y por esta razón, el autor del método, GOWER (56), propuso el nombre de MEDIANA y derivó sus propiedades del teorema de APOLLONIUS.

Aunque este método puede ser apropiado para medidas de similitud y de distancia, LANCE y WILLIAMS (84) indicaron que es incompatible para medidas tales como los coeficientes de correlación de difícil interpretación geométrica.

F.5. METODO DE LA DISTANCIA MEDIA:

Este método define la distancia entre grupos, como la media de las distancias entre todos los pares de individuos de los dos grupos.

SOKAL y MICHENER (115) utilizaron esta media como una medida de distancia entre un individuo y un grupo de individuos, mientras que LANCE y WILLIAMS (84) la extendieron a una medida de distancia entre grupos.

El procedimiento puede ser utilizado con medidas de similitud y de distancia.

LANCE y WILLIAMS (84) sugirieron la siguiente medida d_{ij}, de disimilitud entre los grupos i y j.

$$d_{ij} = \cos\left[\frac{1}{n_i\, n_j} \sum_{i,j} \cos^{-1} S_{ij}\right]$$

siendo n_i y n_j los números de lementos de los grupos y S_{ij} una medida simple entre individuos.

F.6. METODO DE CATTELL:

Este método es muy utilizado en psicología y se denomina también ANALISIS FACTORIAL Q; el tipo corriente de análisis factorial, que opera con correlaciones entre variables, se llama análisis factorial R .

En el presente método, se intercambian individuos y variables con respecto a un "factor de análisis" normal,de modo que las correlaciones pasan a ser correlaciones entre individuos mejor que entre variables.

Los métodos corrientes del análisis factorial incluyen rotación y se aplican a la matriz de correlaciones. Los individuos son, entonces,asignados a los grupos, dependiendo del "peso" del factor.

F.7. CRITERIO DE OPTIMIZACION DE LA COVARIANZA:

Se colocan los puntos en k conglomerados y, luego, se hacen reasignaciones, teniendo en cuenta un criterio de varianza-covarianza.

Si T es la matriz que representa la dispersión total de n puntos en un p-espacio, W_i es la matriz de dispersión de n_i puntos en el conglomerado i y B es la matriz de dispersión de los centroides de K conglomerados, entonces, $T = W + B$, siendo $W = \Sigma W_i$.

Pueden ser desarrollados, entre otros, dos criterios: minimizar la traza de W y maximizar el cociente $|T|/|W|$ siendo

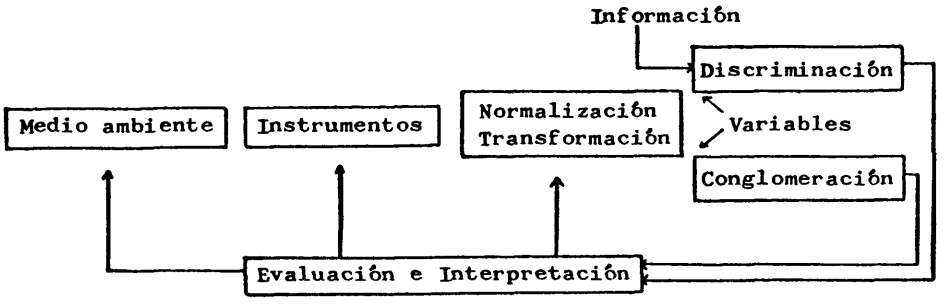
preferido este último criterio por ser invariante bajo transformaciones lineales.

VIII. IDENTIFICACION Y ADSCRIPCION DE NUEVOS OBJETOS O UNIDADES TAXONOMICAS A CLASES YA ESTABLECIDAS:

Como se ha dicho anteriormente, este aspecto corresponde al método estadístico denominado ANALISIS DISCRIMINANTE que es un problema estadístico clásico tratado, entre otros, por FISHER y MAHALANOBIS y que forma parte del RECONOCIMIENTO DE PATRONES.

"El proceso de definir en cual, de un conjunto de grupos definidos a priori, se coloca un individuo, se llama IDENTIFICACION o ASIGNACION", DAGNELIE (22).

Un esquema del proceso clasificatorio puede ser el siguiente:



Pueden ofrecerse otros esquemas clasificatorios en los que,una vez efectuada, según determinados criterios, la conglomeración y fijados los grupos, hemos de colocar nuevos entes en los mencionados grupos. Aquí surge el problema de la asignación.

Un procedimiento sencillo de adscripción, sería incluir el objeto o unidad taxonómica en las clases cuyos promedios de

valores de los caracteres considerados estén más próximos al punto representativo del objeto o unidad taxonómica en el espacio de caracteres. La proximidad supone una previa definición de DISTANCIA entre el centro de gravedad de la clase y el punto a clasificar, distancia que no necesariamente es la euclidea, sino que puede introducir coeficientes de corrección de las posibles correlaciones entre caracteres o tener cualquier expresión analítica que se considere conveniente.

En general, podemos hablar de la situación típica del ANALISIS DISCRIMINANTE: una vez conocida la existencia de dos o más clases, y conocida una muestra de individuos de cada una de ellas, ha de conseguirse un procedimiento o regla mediante la cual, un objeto de origen desconocido pueda ser asignado a una clase correctamente en algún sentido óptimo como puede ser, por ejemplo, la minimización del error.

Es característico del análisis discriminante que las clases dadas sean distintas y separadas, mientras que en el análisis de conglomerados no conocemos de antemano si los individuos del conjunto en escrutinio irán a clases separadas o no.

Por tanto, el problema de la conglomeración es anterior al de la discriminación; sin embargo, estas dos materias se han desarrollado en orden inverso.

El análisis discriminante ha sido estudiado intensamente por los estadísticos durante los últimos cuarenta años, mientras que el análisis de conglomerados es mucho más reciente. La razón, según afirma KENDALL,es que para el uso del análisis de conglomera-

dos se requiere la ayuda de una computadora electrónica que, además, invertirá grandes cantidades de tiempo.

T. CACOULLOS (19), director del Instituto de Estudios Avanzados sobre Análisis Discriminante y sus Aplicaciones, nos dice:

"El Análisis Discriminante, también conocido como Teoría de la clasificación, es una de las más importantes áreas del análisis estadístico. Un problema de discriminación estadística o de clasificación consiste en asignar un individuo o grupo de individuos a una o varias poblaciones conocidas o desconocidas sobre la base de varias medidas de los individuos y muestras de las poblaciones desconocidas".

BALL (11) nos ofrece una tabla en la que clasifica las técnicas de conglomeración y discriminación "de modo que para cada técnica de discriminación es posible encontrar una técnica similar de conglomeración".

PROCEDIMIENTOS DE CLASIFICACION

CONGLOMERACION	DISCRIMINACION
ANALISIS DE FACTORES (RUMMEL, 1967)	ANALISIS DISCRIMINANTE (FISHER, 1936)
FORMACION DE GRUPOS (SOKAL y SNEATH, 1963)	VECINO MAS PROXIMO (COVER y HART, 1967)
PARTICION (FRIEDMAN y RUBIN, 1967)	INVESTIGACION MODAL (ROSEN y MALL, 1966)
DESCOMPOSICION DE MEZCLAS (SPRAGINS, 1966)	TEORIA DE DECISION (CHERNOFF y MOSES, 1959)
METODOS DEL CORREDOR	MAQUINAS LINEALES ADAPTABLES (NILSSON, 1965)

Las técnicas de conglomeración incluyen:

- ENUMERACION TOTAL (FORTIER y SOLOMON, 1966)
- TEORIA DE GRAFOS (ZAHN, 1969)
- FUNCIONES POTENCIALES (MEISEL, 1968)
- TECNICAS DE FUNCIONES ORIENTADAS (GILMAN, 1970)
- APPROACHES DE ACUMULACION (BUTLER, 1969)
- CONGLOMERACION POR LINEAS (EUSEBIO y BALL, 1968)
- CONGLOMERACION INTERACTIVA (HALL, BALL y WOLF, 1969)

Los primeros científicos que utilizaron el ANALISIS DISCRIMINANTE fueron R.A. FISHER, P.C. MAHALANOBIS, C.R. RAO, T.W. ANDERSON, D.G. KENDALL y R.A. REYMENT.

UN NUEVO METODO DE CONGLOMERACION: EL METODO δ

Como hemos visto anteriormente, existen diferentes métodos de conglomeración.

Una vez definidos los OTUS a clasificar y seleccionada una distancia, se inicia el proceso de la conglomeración, que puede llevarse a cabo según diferentes algoritmos, que difieren unos de otros en la distancia elegida entre los conglomerados.

Nosotros proponemos un nuevo algoritmo, al que denominamos y cuya característica fundamental es que tiene en cuenta los valores máximo y mínimo de la distancia entre los OTUS, pretendiendo determinar, al igual que los métodos ya establecidos, la estructura básica, o mejor, la estructura natural de un conjunto de elementos.

En todo método de conglomeración, se elige una distancia d entre los OTUS y se define una distancia d' entre los conglomerados; y es en la definición de d' donde reside la distinción fundamental entre unos y otros métodos.

Nosotros proponemos la siguiente distancia entre conglomerados:

DEFINICION: Sean los conglomerados
$A = \{a_1, a_2 \dots \dots a_h\}$ y $B = \{b_1, b_2 \dots \dots b_k\}$ constituído respectivamente, por los OTUS
$a_i, i = 1 \dots \dots h$ y $b_j, i = 1 \dots \dots k$,

$$\delta(A,B) = \frac{\max\{d(a_i,b_i)\} + \min\{d(a_i,b_i)\}}{2} \qquad \begin{matrix} i = 1,2,\ldots h \\ j = 1,2,\ldots k \end{matrix}$$

donde d es una métrica asociada al conjunto X, inicial, de OTUS.

Tenemos así un procedimiento o algoritmo de conglomeración que tiene en cuenta los valores máximos y mínimos de d para el conjunto X, por lo que, podríamos decir, que este nuevo método es"intermedio" entre los métodos de la distancia máxima y de la distancia mínima, debidos a JOHNSON (1967), como lo es el método de la media ( ), debido a SOKAL y MICHENER (1958), pero difiere esencialmente de este último ya que mientras que estos autores consideran la media aritmética de las distancias máxima y mínima nosotros consideramos la media aritmética de los valores máximo y mínimo de las distancias, d, entre TODOS los elementos de dos conglomerados, como puede verse en la clasificación que hemos hecho según los métodos anteriormente citados de ciertos ejemplares de NEUROPTERIS, y comparando los coeficientes de correlación cofenéticos entre los datos originales y los dendrogramas correspondientes, obtenemos mejores resultados, con nuestro método δ que los que se logran con los algoritmos de la distancia máxima y de la media.

DESCRIPCION DEL METODO .-

1. Sea $X = \{x_1, x_2 \ldots \ldots, x_n\}$ un conjunto finito de OTUS que deseamos clasificar.

A cada uno de ellos se le asocian sus caracteres, es decir, un conjunto de p números reales que nos permiten considerar a los OTUS como puntos de un p-espacio euclideo cuyos p ejes de

de coordenadas son los caracteres.

Sea d, la métrica elegida, es decir, d es una aplicación del conjunto X x X en el conjunto R de los números reales.

$$d: X \times X \longrightarrow R$$

$$d(x_i,x_j) = dij$$

Obtenemos así un conjunto de $\frac{1}{2}$ n (n-1) números reales o distancias dij entre OTUS que disponemos en la matriz simétrica,D de dimensiones n x n:

D_1	x_1	x_2	... ...	x_n
x_1	0	d_{12}	... ...	d_{1n}
x_2		0		d_{2n}
⋮			⋱	
x_{n-1}				$d_{n-1,n}$
x_n				0

que se asocia a la partición inicial P_1: $(x_1),(x_2)$ ... ... (x_n)

2. Sea $d(x_i,x_j)$ = mín dij/i= 1,2,... n, j = 1,2, ... n entonces los OTUS x_i y x_j se fusionan y pasan a formar el conglomerado x_i,x_j al nivel de conglomeración $d(x_i,x_j)$.

La nueva partición es, ahora,

$$P_1: \{x_i\}, \{x_2\} \ \ldots \ \ldots \ \{x_i, x_j\}, \ \ldots \ \{x_n\}$$

Hemos de obtener la correspondiente matriz de distancias D_1 asociada a P_1, que será una matriz simétrica de dimensiones (n-1) x (n-1).

Las distancias entre los conglomerados constituídos por un único elemento serán las mismas que las que figuran en la matriz D_1, y para obtener las"nuevas" distancias entre el conglomerado (x_i, x_j) y los restantes elementos de P_1 consideramos la distancia δ , que define y caracteriza este método, de la siguiente forma:

$$\delta[\{x_i, x_j\}, x_k] = \frac{\text{máx}\left[d(x_i, x_k),\ d(x_j, x_k)\right] + \text{mín}\left[d(x_i, x_k),\ d(x_j, x_k)\right]}{2}$$

La matriz D_2 asociada a esta participación será,ahora:

D_2	$\{(x_i, x_j)\}$	$\{x_1\}$	$\{x_2\}$	... ...	$\{x_n\}$
$\{x_i, x_j\}$	0	δ_{11}	δ_{12}	... ...	δ_{1n}
$\{x_1\}$		0	d_{23}	... ...	d_{2n}
$\{x_2\}$			0	... ...	d_{3n}
⋮		⋮	⋱	⋱	⋮
$\{x_n\}$					0

3. Se elige de nuevo, el menor elemento de la matriz anterior y los correspondientes elementos pasan a formar un nuevo conglomerado, al nivel correpondiente, lo que da lugar a una nueva partición para la que, aplicando la definición de la distancia δ entre conglomerados, se calcula su matriz asociada de distancias.

4. Se repite el proceso hasta que todos los elementos hayan sido agrupados a un determinado nivel.

5. Se construye el correspondiente dendrograma en el que se reflejarán los niveles que, en cada fusión, se han ido obteniendo.

En el capítulo V se ofrece la clasificación que hemos hecho, aplicando el método δ , de 13 ejemplares de Neuropteris y una comparación de los resultados con los obtenidos mediante otros algoritmos.

1* : Así como el término homología está relacionado con la semejanza global, existe otro término, analogía que se utiliza para expresar la semejanza intermedia de una relación continua en la que los extremos serían la identidad total y la disimilitud total. Según SNEATH y SOKAL (112): "mejor que hacer un contraste categórico entre homología y analogía, es preferible hablar de mayor o menor semejanza entre órganos, sistemas de órganos y otros niveles de complejidad". Pero, existen, principalmente, en Zoología y Botánica otras acepciones de los términos homología y analogía.
STORER y USINGER en su obra "GENERAL ZOOLOGY"(Mac Graw-Hill (1957) hablan de analogía al referirse a la similitud en la función y no en el origen, por ejemplo, alas de insectos y alas de vertebrados y reservar el término homología para reflejar la similitud en el origen y no en la función.
Son, por ejemplo, órganos homólogos los pétalos de las rosas, los zarcillos de los guisantes y las espinas de los agracejos. Véase, también "DICCIONARIO DE BOTANICA de P. FONT QUER, Edit. Labor, 1975".

2* : Los "estados" de un carácter también se llaman modalidades.

3* : Cuando los caracteres son cualitativos se utiliza la palabra "atributo".

4* : W.H. KRUSKAL y J.M. TANUR en su otra "INTERNATIONAL ENCYCLOPEDIA OF STATISTICS" hablan de métodos bilineales en Análisis Plurivariante para designar aquellos métodos que se basan en la consideración de matrices en las que las filas son unidades taxonómicas operacionales y las columnas representan caracteres.

5* : En muestreo de poblaciones finitas se llama conglomerado a cualquier unidad de muestreo constituída por varias unidades últimas o de objetos en estudio.

6* : Estos puntos se denominan, a veces, PATRONES (PATTERN)porque son una representación de un individuo u OTU definido por los valores o estados de los caracteres en consideración.

7* : Es importante observar que en los problemas de muestreo de poblaciones finitas, los conglomerados, para que realmente sean representativos, han de ser heterogéneos.

8* : La entropía o información no es una medida de la "cantidad" de conocimiento que se posee sobre un grupo, sino, más bien

es una medida del desorden de la variación de la confusión (112).
Hay que distinguir entre entropía intrataxónica, que permite clasificar elementos y entropía intertaxónica que se utiliza para la evaluación de clasificaciones.
J. GARRIDO y F. AZORIN en "ENTROPIA Y CLASIFICACION DE LAS ESTRUCTURAS CRISTALINAS", U.A.M. Madrid (1976) relacionan el concepto de especificidad de los caracteres taxonómicos con la valoración de la calidad y utilidad de las diversas clasificaciones de las estructuras cristalinas y discuten esta noción de calidad aplicando el concepto de entropía de una clasificación que mide la homogeneidad interna de los grupos taxonómicos y su relación con las diferencias existentes entre los diversos grupos.
Consideran, especialmente, el caso de los caracteres cualitativos aplicándolo a las clasificaciones mineralógicas.

CAPITULO II

ESTUDIO DE LA ANALOGIA TAXONOMICA

INTRODUCCION

MEDIDAS DE ASOCIACION

COEFICIENTES ANGULARES

MEDIDAS DE DISTANCIA

I. METRICA DE MINKOWSKI:
II. METRICA DE CANBERRA:
III. DISTANCIA ABSOLUTA:
IV. DISTANCIA DE MAHALANOBIS:
V. DISTANCIA GENERAL CON PESOS:
VI. COEFICIENTE DE DIVERGENCIA:
VII. DISTANCIA DE JEFFREYS-MATUSITA:
VIII. COEFICIENTE DE PARECIDO RACIAL:
IX. COEFICIENTE DE ROGERS Y TANIMOTO:
X. IVANOVIC:
XI. DIFERENCIAS EN TAMAÑO Y FORMA:
XII. COEFICIENTES DE DISIMILITUD QUE MIDEN LA VARIACION DENTRO DE LOS OTUS:

MEDIDAS PROBABILISTICAS

MEDIDAS FUNCIONALES DE DISIMILITUD

OTROS COEFICIENTES DE DISIMILITUD

INTRODUCCION

El problema de la analogía taxonómica consiste en encontrar el mayor número de relaciones entre los elementos de un conjunto dado, bien sea dos a dos, bien entre cada uno de ellos y los demás.

En esta definición encontramos dos palabras clave, elemento y relación, que son elementos fundamentales del problema taxonómico.

Los objetos, que también se llaman elementos o unidades taxonómicas operacionales, "OTUS" [1*] constituyen la entidad básica del problema de clasificación, tanto en el sentido de formación de clases como en el de asignación, o discriminación y están definidos por un conjunto de medidas, estados o números, correspondientes a los caracteres elegidos para identificarlos

Tales variables [2*] caracterizan el objeto y contienen toda la información acerca del mismo, por tanto, cada objeto u OTU estará descrito mediante las variables o caracteres.

Una vez conocidos los OTUS habrá que seleccionar las variables, procurando reducir su número, eliminando las menos efectivas, lo que es equivalente a reducir la dimensionalidad del espacio de medida, que tendrá tantas dimensiones como caracteres los OTUS, tratando de mantener, siempre, la estructura de las observaciones.

Las variables o caracteres pueden ser numéricas, nominales, ordinales, binarias, multifásicas o multivariantes y condicionalmente presentes.

Una vez que se conocen los valores o estados de los caracteres de todos los objetos u OTUS a estudiar, se disponen u ordenan en una MATRIZ DE DATOS ORIGINALES [3*] de n filas, correspondientes a n elementos y p columnas, correspondientes a p caracteres, es decir, cada elementos de la misma X_{ij} es el resultado de la observación del carácter j en el OTU i; las p columnas representas los p caracteres y las n filas, los n OTUS.

ELEMENTOS o UTOS	CARACTERES				
	1	2	3		p
2	x_{11}	x_{12}	....		x_{1p}
3	x_{21}	x_{22}	....		x_{2p}
⋮	⋮				⋮
n	x_{n1}	x_{n2}	....		x_{np}

Antes de calcular los coeficientes de analogía se suele aplicar a la matriz de datos originales alguna transformación tendente a minimizar o eliminar propiedades indeseables para el cálculo posterior y a facilitar el mismo; estas transformaciones usualmente, son transformaciones lineales y tipificaciones o el cálculo de las denominadas rankitas (rankits).

El estudio de la matriz de datos puede hacerse utilizando dos grandes grupos de técnicas, correspondientes a dos distintos puntos de vista:

- La técnica Q estudia la asociación de los OTUS (filas) en base a los caracteres (columnas)

- La asociación de pares de los p caracteres, columnas, puede ser examinada sobre todos los OTUS (filas). Es la llamada técnica R.

Una vez determinada la matriz de datos transformados o no, ya podemos acometer el estudio de la analogía entre todas las parejas de OTUS calculando coeficientes de analogía que no son otra cosa que herramientas matemáticas que cuantifican el parecido de los objetos.

Los conceptos generales de analogía, semejanza y parecido, y sus complementarios se concretan, desde ahora en otros dos que utilizaremos constantemente: SIMILITUD y DISIMILITUD.

"Los métodos basados en la idea de similitud o disimilitud constituyen, como dice SIBSON (109) el más importante grupo de técnicas para analizar matrices de objetos por caracteres".

Antes de pasar al apartado siguiente, queremos insistir en la gran importancia que tiene el conocimiento de la métrica del espacio en el que se definen los OTUS con objeto de conocer e interpretar las relaciones fenéticas que existen entre los mismos.

Los diferentes coeficientes o medidas a utilizar, los vamos a dividir en cinco grupos:

- MEDIDAS DE ASOCIACION
- COEFICIENTES ANGULARES
- MEDIDAS DE DISTANCIA
- MEDIDAS PROBABILISTICAS
- MEDIDAS FUNCIONALES DE SIMILITUD

Todas las medidas o funciones de semejanza son aplicaciones del conjunto producto X x X en el conjunto $\mathbb{R}$, siendo X el conjunto o espacio de los objetos, individuos, elementos u OTUS

a clasificar y que tendrá un número de dimensiones igual al de caracteres o atributos en consideración.

A cada OTU; X_i se le asocia un vector $(x_{i1}, x_{i2}, \ldots x_{ip})$ cuyas coordenadas corresponden al valor del carácter j para dicho OTU.

De esta forma podemos considerar un espacio de vectores E y toda medida de similitud S, como una aplicación de $E \times E \xrightarrow{S} \mathbb{R}$.

Hay que observar que en muchos campos de estudio, las entidades u OTUS son individuos únicos y, en otros, como, por ejemplo en la taxonomía de organismos vivos, los OTUS son poblaciones bien definidas y las observaciones tendrán, pues, una distribución de probabilidad.

En todo caso, cuando todas las coordenadas de los vectores son cuantitativas, es natural considerar el conjunto E como n puntos en un espacio de dimensión p (número de caracteres) que suele ser geométrico o euclideo, aunque ésto no es una condición necesaria.

HARTIGAN da la siguiente lista de doce estructuras de similitud:

1.- S definida sobre E x E es la distancia euclidea.

2.- S definida sobre E x E es una métrica

3.- S definida sobre E x E es simétrica con valores reales.

4.- S definida sobre E x E toma valores reales.

5.- S es un orden completo, $\preceq$, sobre E x E.

6.- S es un orden parcial, $\leq$, sobre E x E (cada par comparable de entidades, puede ser ordenado, pero no todos los pares de objetos son comparables).

7.- S es un árbol $\mathcal{T}$ sobre E un orden parcial de similitud, $(i,j) \leq (k,l)$ siempre que

$$\sup_{\mathcal{T}} (i,j) \geq \sup \ (k,l)$$

8.- S es un orden completo relativo de similitud $\leq i$, sobre E, para cada entidad X_i de E: $j \leq_i k$ significa que X_j no es más similar a X_i que X_k.

9.- S es un orden parcial relativo de similitud $\leq_i$, sobre E.

10.- S es una dicotomía de similitud sobre E x E en donde E x E está dividido en un conjunto de pares similares y un conjunto de pares no similares.

11.- S es una tricotomía de similitud sobre E x E (pares similares, pares no similares y el resto).

12.- S es una partición de E en conjuntos de objetos similares.

MEDIDAS DE ASOCIACION

Estas medidas, que tienen una larga historia, se utilizaron, en principio, en taxonomía biológica y tienen por objeto medir la similitud o afinidad entre dos OTUS a partir de los valores de un conjunto de caracteres comunes a ambos.

Los coeficientes de asociación más simples, expresan la razón entre las identidades o coincidencias observadas en los estados de los caracteres para el par de OTUS en cuestión y el número total de caracteres.

Estos coeficientes suelen tomar valores comprendidos entre 0 y 1, característica que los diferencia de las distancias que pueden adoptar cualquier valor, cuando no están previamente estandarizadas.

En la mayoría de los casos en los que se utilizan estas medidas, las variables son del tipo PRESENCIA-AUSENCIA que se suelen notar por 1(+) ó 0(-).

Los atributos cuantitativos pueden utilizarse únicamente en aquellos casos en que se pueda encontrar alguna forma o procedimiento de "partir" (dividir) el recorrido de variación en intervalos tales que dicho recorrido de variación de cada población esté comprendido enteramente dentro de un intervalo único.

Los datos cualitativos, para lo que las variables pueden tomar diferentes estados, modalidades o niveles, se deben tratar de forma similar a los datos binarios con cada nivel de variación, que se considerará como una variable binaria y única.

Es muy importante observar que los distintos coeficientes de similitud pueden tomar diferentes valores para el mismo conjunto de datos, lo que es fácil de comprender, puesto que cada uno de ellos expresa un aspecto distinto de la similitud entre los objetos.

SOKAL y SNEATH (115) discutieron el uso de los coeficientes de similitud para datos binarios y decidieron que no se puede establecer ninguna regla para la utilización de uno u otro coeficiente, sino para cada conjunto de datos debe ser considerado "por sus propios méritos" por aquel investigador más familiarizado con dicho material.

Estos coeficientes se calculan sobre las columnas de una matriz y las fórmulas pueden generalizarse para las matrices aleatorias.

DEFINICION: Una función con valores reales no negativos

$$S(X_i, X_j) = S_{ij}$$

se dice que es una medida de similitud si,

1. $0 \leq S(X_i, X_j) \leq 1$ para $X_i \neq X_j$
2. $S(X_i, X_j) = 1$ si y solo si $i = j$
3. $S(X_i, X_j) = S(X_j, X_i)$

A partir de las similitudes se construye la matriz de similitud:

$$\begin{matrix} 1 & S_{12} & \cdots & \cdots & \cdots & S_{1n} \\ S_{21} & 1 & \cdots & \cdots & \cdots & S_{2n} \\ \vdots & & & & & \\ S_{n1} & & \cdots & \cdots & \cdots & 1 \end{matrix}$$

en la que cada elemento es un coeficiente de similitud o coeficiente de asociación.

Supongamos que cada vector observación contiene únicamente ceros y unos, es decir, tenemos datos binarios; dados dos vectores de medida X_i y X_j definimos:

a. M_{IJ} como el número de características que toman el valor 1 en X_i y en X_j.

b. M_{ij} es el número de veces que toma el valor cero en ambos.

c. M_{iJ} el número de veces que toma 0 en X_i y 1 en X_j.

De esta forma: $M_J = M_{IJ} + M_{iJ}$ es el número de unos en X_j y $M_j = M_{Ij} + M_{ij}$ es el número de ceros en X_j.

A continuación, damos la tabla de coeficientes de SIMILITUD en términos de las cantidades anteriormente definidas.

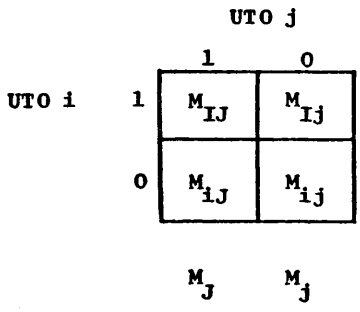

UTO j
1
0
UTO i
1
0
M_{IJ}
M_{Ij}
M_{iJ}
M_{ij}
M_J
M_j

TABLA DE COEFICIENTES DE SIMILITUD PARA DATOS BINARIOS:

$$\frac{M_{IJ}}{IJ + M_{IJ} + M_{ij}}$$ JACCARD (1908), SNEATH (1957)

$$\frac{M_{IJ} + M_{ij}}{P}$$ SOKAL y MICHENER (1958)

$$\frac{M_{IJ}}{P}$$ RUSSEL y C.R. RAO (1940)

$$\frac{2M_{IJ}}{2M_{IJ} + M_{IJ} + M_{iJ}}$$ SORENSEN (1948), DICE (1845)

$$\frac{2\,(M_{IJ}+M_{ij})}{P + M_{IJ} + M_{ij}}$$

$$\frac{M_{IJ}}{M_{IJ} + 2(M_{IJ} + M_{ij})}$$

$$\frac{M_{IJ} + M_{ij}}{P + M_{Ij} + M_{iJ}}$$ ROGERS y TANIMOTO (1960).

$$\frac{M_{IJ}}{M_I + M_J - 2M_{IJ}}$$ KULCZYNSKI (1927)

$$\frac{M_J}{M_j}$$

$$\frac{1}{2}\frac{M_{IJ}}{M_I} + \frac{M_{IJ}}{M_J}$$ KULCZYNSKY (1927)

$$\frac{1}{4}\frac{M_{IJ}}{M_I} + \frac{M_{IJ}}{M_J} + \frac{M_{ij}}{M_j} + \frac{M_{ij}}{M_j}$$

$$\frac{M_{IJ}}{M_I M_J}$$ OCHIAL (1957)

$$\frac{M_{IJ} M_{ij}}{M_I M_J M_i M_j}$$

$$\frac{M_J - M_I}{M_J + M_j}$$ HAMANN (1961)

$$\frac{M_{IJ}\,M_{ij} - M_{Ij}\,M_{iJ}}{M_{IJ}\,M_{ij} + M_{Ij}\,M_{iJ}}$$

YULE (1911)

$$\frac{M_{IJ}\,M_{ij} - M_{iJ}\,M_{Ij}}{(M_I\,M_J\,M_i\,M_j)^{\frac{1}{2}}}$$

PEARSON

Pueden definirse nuevos coeficientes de asociación como promedios simples o ponderados de los anteriormente definidos.

Finalmente vamos a hacer un breve comentario sobre el COEFICIENTE GENERAL DE SIMILITUD DE GOWER (1971), de gran trascendencia en taxonomía y del que son casos particulares la mayoría de los coeficientes de similitud. Puede ser utilizado para datos binarios, cualitativos y cuantitativos.

Su forma es : $$S_{ij} = \sum_{i=1}^{p} W_{ijk} \; S_{ijk} \; / \sum_{i=1}^{n} W_{ijk}$$

Los pesos, W_{ijk}, toman los valores 1 ó 0 dependiendo de que la comparación se considere válida para la variable k, o que dicha variable sea desconocida para uno al menos de los UTOS en estudio, exceptuando el caso de variables dicotómicas, en el que W_{ijk} se toma igual a cero, cuando la mencionada variable no figura en los UTOS. Siempre que $W_{ijk} = 0$, S_{ijk} es igual a cero y si W_{ijk} es cero para todas las variables, entonces S_{ij} no está definido.

Los valores de S_{ijk} $0 \leq S_{ijk} \leq 1$ se asignan de la forma siguiente:

a. DATOS BINARIOS:

UTO	i	+	+	-	-
UTO	j	+	-	+	-
S_{ijk}		1	0	0	0
W_{ijk}		1	1	1	0

b. DATOS CUALITATIVOS:

En este caso $S_{ijk} = 1$ si los UTOS i y j son iguales para el caracter k y, $S_{ijk} = 0$ si difieren.

c. DATOS CUANTITATIVOS:

En este caso, $S_{ijk} = 1 - \frac{|X_{ik} - X_{jk}|}{R_k}$

donde X_{ik} es el valor del UTO i para la variable k y R_k es el rango de la variable k.

El coeficiente de similitud de GOWER se puede utilizar, también, con datos que contengan mezcla de variables binarias cualitativas y cuantitativas.

Entre las medidas de asociación se pueden considerar las medidas entre variables nominales, que ofrecemos a continuación:

En la siguiente tabla de contingencia representamos la distribución de dos variables, ordinales o nominales). El elemento n_{ij} es el número de coincidencias de la variable A en la clase i y de la variable B en la clase j.

Variable A	Variable B	1	2	...	...	q	Totales
1		n_{11}	n_{12}	...	...	n_{1q}	n_1
2		n_{21}	...	...	...	n_{2q}	n_2
:		:					
p		n_{pl}	...	...	...	n_{pq}	n_p
Totales		n.1	n.2	...	...	n.q	n..

Si todos los elementos de la tabla junto con los marginales totales se dividen por n.., que es el número total de elementos, expresaríamos los elementos en términos de frecuencias, - fij. Las medidas ordinales de asociación dependen de la sucesión de filas y columnas mientras que las medidas nominales son invariantes a cualquier permutación de filas y columnas.

Veamos los siguientes tipos de medidas entre variables nominales:

a. Medidas basadas en la X^2

Si D_{ij} es el valor observado en la célula ij y e_{ij} es el correspondiente valor esperado bajo la hipótesis de independencia, el estadístico muestral X^2 se define como

$$X^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} (D_{ij} - e_{ij})^2 / e_{ij}$$

El valor de X^2 aumenta cuando lo hace n.. y no es una buena medida de asociación. Un remedio parcial está en

$$\phi^2 = X^2 / n..$$

que se denomina la contingencia media cuadrática, cantidad que depende del tamaño de la tabla; se han hecho varios intentos - para que ϕ^2 varie entre 0 y 1:

TSCHUPROW propuso la media geométrica de p-1 y q-1 como factor normalizante y obtuvo la siguiente medida

$$T = \left(\frac{X^2/n..}{(p-1)\ (q-1)^{\frac{1}{2}}} \right)^{\frac{1}{2}}$$

CRAMER dio la medida:

$$C = \left(\frac{X^2/n..}{\text{min. } (p-1),(q-1)} \right)^{\frac{1}{2}}$$

PEARSON sugirió otra medida basada en ϕ^2:

$$P = \left(\frac{\phi^2}{1+\phi^2} \right)^{\frac{1}{2}} = \left(\frac{X^2}{n..+X^2} \right)^{\frac{1}{2}}$$

medida que se llama COEFICIENTE DE CONTINGENCIA

b. Medidas basadas en la predicción de clase óptima.

Una de las formas más atractivas de dar significado a la asociación entre variables es midiendo el poder de una variable como predictora de la otra.

Si suponemos que el objetivo es "adivinar" la clase B para una observación.

Si las observaciones de la clase A son desconocidas, la mejor forma de hacerlo es elegir la clase B con la mayor frecuencia o probabilidad marginal total, es decir, el valor de m que satisface:

$$f.m = \text{máx } f._1, \ldots, f._q$$

"La probabilidad de error en este caso es $P_1 = 1 - f.m$ "

Supongamos ahora que es conocido a, el valor de una observación de la clase A. Entonces solo interesa la fila a de la tabla de contingencia y la conjetura es la clase B que corresponda al mayor elemento en la fila a, es decir, el valor de m_a que satisface

$$f_{am_a} = \max \{f_{a1}, \ldots \ldots, f_{aq}\}$$

donde el subíndice a de m indica que la clase óptima B puede variar de fila a fila.

Como a es conocido la probabilidad de error es $P_1 = 1 - F_{am_a}/f_a$. Pero hay p filas, cada una de las cuales tiene frecuencia f_i.

Por tanto, la probabilidad incondicional de error es

$$P_2 = \sum_{i=1}^{P} f_i(1 - f_{im_i} / f_i) = 1 - \sum_{i=1}^{P} f_{im_i}$$

GOODMAN y KRUSKAL sugirieron una medida del poder predictivo de A sobre B

$$L_B = \frac{P_1 - P_2}{P_1} = \frac{\sum_{i=1}^{P} f_{im_i} - f_{\cdot m}}{1 - f_{\cdot m}}$$

Análogamente, la predicción de la clase A está dada por:

$$L_A = \frac{\sum_{j=1}^{q} f_{m_j j} - f_{m \cdot}}{1 - f_{m \cdot}}$$

Otra medida relacionada con L es

$$D = P_1 - P_2 = \tfrac{1}{2}\left[\sum_{i=1}^{P} f_{im_i} + \sum_{j=1}^{q} f_{m_j j} - f_{m} - f_{m}\right]$$

ó

$$D = \frac{\sum_{i=1}^{P} M_i m_i + \sum_{j=1}^{q} Mm_j j - M.m - Mm.}{2n\ ..}$$

Y terminamos este apartado con la siguiente frase de JARDINE y SIBSON: "La elección de una medida de asociación entre individuos, suele estar determinada, frecuentemente, por la naturaleza del problema clasificatorio".

COEFICIENTES ANGULARES

A. La utilización del ángulo como una medida del parecido o afinidad entre UTOS fue sugerido, en primer lugar por BHATTACHARYYA (35).

Con posterioridad, EDWARS y CAVALLI-SFORZA (1964), en un estudio de relaciones entre poblaciones humanas, midió el parecido o afinidad por ángulos entre puntos, que representaban las poblaciones sobre la superficie de una hiperesfera unidad.

El coseno del ángulo que forman los vectores que unen los puntos, imágenes geométricas de los UTOS, con el origen de coordenadas está dado por la fórmula:

$$\text{Cos.} \propto = \frac{\sum_{i=1}^{M} X_{ij} X_{ik}}{\sqrt{\sum_{i=1}^{M} X_{ij}^{2}} \sqrt{\sum_{i=1}^{M} X_{ik}^{2}}}$$

Cuando los caracteres utilizados al hacer una comparación, son principalmente morfológicos, a los taxónomos les interesa separar en el estudio de la afinidad entre modelos, las componentes de tamaño y de forma.

Una medida del tamaño nos la puede dar el módulo del vector cuyo origen es el de coordenadas y su extremo, el punto que representa una UTO.

Dos puntos a diferentes distancias, pero alineados con

el origen, representan dos modelos que difieren únicamente en el tamaño y dos puntos no alineados con el origen, pero a igual distancia del mismo, representan dos modelos que difieren únicamente en la forma.

Dos modelos pertenecerán a la misma linea, que es incidente con el origen, si sus valores de caracteres (coordenadas) son proporcionales.

Esta clase de diferencias de tamaño es una diferencia PROPORCIONAL y es evidente que si dos modelos pertenecen a la misma linea que pasa por el origen, y así, difieren únicamente en tamaño, el coseno del ángulo entre los correspondientes vectores es 1, lo que indica que la SIMILITUD ES TOTAL.

"El coseno puede interpretarse como una medida de la similitud que ignora diferencias proporcionales de tamaño".

El término DIFERENCIA DE TAMAÑO puede utilizarse para describir la situación en que los valores de los caracteres de un modelo o UTO difieren de los de otro en alguna cantidad constante. Es ésta una diferencia de tamaño ADITIVA.

B. El segundo coeficiente angular que vamos a considerar es el COEFICIENTE DE CORRELACION que no tiene en cuenta las diferencias de tamaño aditivas y proporcionales y que es igual al coseno del ángulo de los vectores cuyas coordenadas son los valores de los caracteres de los UTOS expresados como desviaciones respecto de la media de todos los valores de caracteres de cada UTO.

Como indicó MINKOFF (112) el uso taxonómico del coeficiente de correlación como una medida de similitud exige que todos los caracteres tengan las mismas propiedades direccionales y dimensionales.

La fórmula del coeficiente de correlación es:

$$S_{jk} = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2} \sqrt{\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2}}$$

para los UTOS j y k, siendo x_{ij} el valor del caracter i en la UTO j, $\bar{x}_j$ la media de todos los valores de los caracteres para la UTO j, y n el número de caracteres.

Como la fórmula anterior se basa en momentos respecto de la media, considera la no asociaciones entre las UTOS para caracteres con más de dos estados, por lo que, en este aspecto, los coeficientes de correlación son más eficaces que los coeficientes de asociación.

C. COEFICIENTE DE PARECIDO RACIAL DE K. PEARSON (1926) ("Coefficient of racial likeness). En la mayoría de las antiguas aplicaciones en Biología y Psicología se utilizaron los objetos individuales como UTOS, y, también, en muchas aplicaciones del análisis de conglomerados, éste es aún el modelo apropiado. Sin embargo, las primeras aplicaciones en Antropología, se basaron en muestras tales que cada UTO (una muestra de una población) no solo se representó por un vector, sino, también, por una matriz de varianzas - covarianzas.

Un primer intento de estudio de estos problemas lo llevó a cabo K. PEARSON con su COEFICIENTE DE PARECIDO RACIAL que permitía las varianzas pero no las covarianzas entre caracteres como distancias estimadas entre muestras de población. La fórmula para este coeficiente es:

$$C.R.L. = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{(\bar{X}_{iJ} - \bar{X}_{iK})^2}{(S^2_{iJ/nJ}) + (S^2_{iK/nK})} \right]^{\frac{1}{2}} - \frac{2}{n}$$

en donde $\bar{X}_{ij}$ representa la media de la muestra del carácter i-ésimo para la muestra J, S^2_{iJ} es la varianza de la misma y n_J, el tamaño de la muestra J.

D. COEFICIENTE DE CORRELACION DE CATTELL. Los coeficientes de distancia basados en caracteres estandarizados pueden transformarse fácilmente en coeficientes de correlación, como ocurre con el coeficiente de CATTELL

$$r_{P(jk)} = \frac{2X^2_{5[nJ} - nd^2_{jk}}{2X^2_{5.n} + nd^2_{jk}}$$

donde $X^2_{.5\ n}$ es el valor medio X^2 para n grados de libertad.

MEDIDAS DE DISTANCIA

Aunque históricamente los coeficientes de asociación o de similitud, incluidos los coeficientes angulares y los de correlación, ocupan el primer lugar en taxonomía, la mayoría de los coeficientes utilizados, sobre todo en análisis de conglomerados, son los coeficientes o medidas de distancia y de disimilitud.

La distancia y la disimilitud pueden transformarse en medidas de similitud mediante diversas transformaciones, como puede ser $S_{ij} = \frac{1}{1+dij}$ o $S_{ij} = c-dij$ siendo c una constante y donde S_{ij} o dij expresan la similitud o disimilitud, respectivamente, entre los OTUS i y j. Sin embargo, pasar de similitudes, ij, a distancias, dij, es mucho más difícil, puesto que las distancias han de satisfacer la desigualdad triangular y es aquí donde reside la diferencia esencial entre estos dos tipos de medidas.

La diferencia más obvia entre similitud y distancia estriba en que las similitudes toman valores comprendidos entre 0 y 1 y las medidas de distancia pueden tomar cualquier valor positivo.

Antes de continuar, hemos de observar que, aunque los términos distancia y disimilitud se utilizan, en general, como si fueran sinónimos, porque expresan (o se refieren a) la desemejanza o disparidad entre objetos u OTUS, desde un punto de vista eminentemente científico, se distinguen perfectamente.

DEFINICION: Sea X un conjunto no vacio de OTUS; una métrica o distancia definida en X es una aplicación de $X \times X \longrightarrow R$, en la que a cada par ordenado (x,y) de elementos de X le corresponde un número real, que

cumple las condiciones siguientes:

Axioma 1.- $d(x,y) \geq 0$, $\forall\, x,y \in X$

Axioma 2.- $d(x,y) = 0$ si y solo si $x = y$

Axioma 3.- $d(x,y) = d(y,x)$, $\forall\, x,y \in X$

Axioma 4.- $d(x,z) \leq d(x,y) + d(y,z)$ $\forall x,y,z \in E$

La condición e exige que la métrica o distancia d sea SIMETRICA respecto de x e y; y la condición 4, que es análoga a la propiedad de las longitudes de los lados de un triángulo, se denomina DESIGUALDAD TRIANGULAR.

DEFINICION: Espacio métrico es el par $\{X,d\}$ formado por un conjunto X, no vacío, y una métrica o distancia definida en el mismo.

De esta definición resulta que dos espacios métricos son distintos cuando difieren en el conjunto soporte X, o cuando teniendo el mismo conjunto X, difieren en las métricas o distancias.

Algunas propiedades de las métricas o distancias y que utilizaremos en este trabajo, son las siguientes:

El axioma 2 implica que la distancia entre un punto y el mismo es cero y, si dos puntos tienen distancia cero, son idénticos.

El axioma 1 prohibe las distancias negativas.

El axioma 3 impone la simetría.

Estas propiedades concuerdan con las nociones intuitivas ya que la concepción popular de una distancia es la distancia euclidea de la geometría elemental, que es una métrica.

"La suma de dos métricas es una métrica".

"El producto de dos métricas(en particular el cuadrado de una métrica) no satisface necesariamente la desigualdad triangular y no es una métrica".

"Cualquier múltiplo positivo de una métrica es una métrica".

"Si d es una métrica y W es un número positivo, entonces $d' = \frac{d}{w + d}$ es también, una métrica".

Si $\{X,d\}$ es un espacio métrico y $X_1 \neq \emptyset$ es un subconjunto de X, considerando la restricción de d a los pares de elementos de X_1, se obtiene una distancia d_1, definida en X_1, pues continúan verificándose las cuatro condiciones que definen una distancia; por tanto "el par $\{X_1, d_1\}$ es un subespacio métrico del espacio métrico $\{X,d\}$ ".

DEFINICION: Una PSEUDOMETRICA[6] en un conjunto X es una función

$d: X \times X \longrightarrow R \cup \{\infty\}$ que verifica las siguientes condiciones:

Axioma 1.- $d(x,x) = 0$, $\forall\, x \in X$

Axioma 2.- $d(x,y) \leq d(x,z) + d(y,z)$.

DEFINICION: Se llama espacio pseudométrico (o semimétrico) al par (X,d).

DEFINICION: Sea X un conjunto no vacío con elementos - llamados OTUS, un COEFICIENTE DE DISIMILITUD sobre X - es una función $d = X \times X \longrightarrow R$ que cumple las siguientes condiciones:

Axioma 1.- $d(x,y) \geq 0 \quad \forall x,y \in X$

Axioma 2.- $d(x,x) = 0 \quad \forall x \in X$

Axioma 3.- $d(x,y) = d(y,x) \quad \forall x,y \in X$

DEFINICION: Un coeficiente de disimilitud se llama coeficiente métrico o métrica o distancia, si verifica la siguiente condición, llamada desigualdad triangular:

$d(x,z) \leq d(x,y) + d(y,z) \quad \forall x,y,z \in X$

DEFINICION: Un coeficiente de disimilitud se llama ultramétrico si cumple la siguiente condición, llamada desigualdad ultramétrica,

$d(x,z) \leq \text{máx} \left\{ d(x,y), d(y,z) \right\}$

Es evidente que esta última condición es mucho más fuerte que la desigualdad métrica, y tiene, además, una gran impotancia.

El axioma 3 de la definición de coeficiente de disimilitud excluye las medidas aritméticas de disimilitud y junto al axioma 2, implica que se pueden especificar completamente las disimilitudes entre los elementos, mediante una matriz triangular de $\frac{n\,(n-1)}{2}$ elementos.

I. METRICA DE MINKOWSKI:

Sea X_{ij} el valor del elemento u OTU j para la variable i, y sea $X_j^T = (x_{ij},\ldots,x_{nj})$ el vector correspondiente al elemento j, entonces la métrica o distancia de MINKOWSKI entre los elementos j y k está dada por:

$$1.\quad d_r\,(X_j X_k) = \sum_{i=1}^{n} \left[|x_{ij} - x_{ik}|^r\right]^{1/r}$$

siendo $r \geq 1$.

Para diversos valores de r obtenemos diferentes distancias:

Si $r = 1$, se obtiene la métrica L_1, de manzanas (city-block) o distancia de MANHATTAN:

$$2.\quad d_1\,(X_j, X_k) = \sum_{i=1}^{n} |x_{ij} - x_{ik}|$$

Si $r = 2$ se obtiene la distancia euclidea o métrica L_2:

$$3.\quad d_2\,(X_j, X_k) = \sum_{i=1}^{n} \left[(x_{ik} - x_{ik})^2\right]^{\frac{1}{2}}$$

Si $r \longrightarrow \infty$ se obtiene la métrica L_∞ o distancia de - CHEBYSHEV:

$$4.\quad d_\infty (X_j,X_k) = \max_{i=1,\ldots,n} |x_{ij}-x_{ik}|$$

que es sencilla computacionalmente, aunque implica un procedimeinto de ordenación.

La distancia de Manhattan, dividida por n, número de -- elementos, nos proporciona la métrica "diferencia media de caracteres de CAIN y HARRISON, o M.C.D.:

$$5.\ \text{M.C.D.}\ (X_j,X_k) = \frac{d_1(X_j,X_k)}{n}$$

que es una medida de gran sencillez estadística y de fácil manejo, aunque infraestima la distancia euclidea entre los OTUS.

Como la distancia euclidea aumenta con el número de caracteres utilizados en la comparación de los OTUS, generalmente - se calcula, a partir de ella, una distancia media, llamada DISTANCIA TAXONOMICA:

$$6.\ d(X_j,X_k) = \sqrt{\frac{d_1^2\ (X_j,X_k)}{n}}$$

Todas las distintas variedades de la métrica de Minkowski tienen tres características:

a. Tanto las variables, como su forma de representación se toman tal y como vienen dadas. Si, por ejemplo, - una variable se expresa en metros y otra en litros -

entonces la métrica incluye la suma de la potencia r de una diferencia en litros.[4*]

b. Cada variable se expresa linealmente, es decir, sin transformación alguna.

c. Cada variable se trata independientemente de las restantes variables. La contribución de cada variable es la potencia r de la diferencia de los datos unitarios y esta cantidad no depende de las otras.

II. METRICA DE CANBERRA:

Ha sido definida por la escuela australiana y tiene por fórmula:

$$7.\quad d(X_j, X_k) = \sum_{i=1}^{n} \left(\frac{|x_{ik} - x_{ik}|}{(x_{ij} + x_{ik})} \right)$$

También ha sido propuesta por LANCE y WILLIAMS.

El numerador es la métrica L_1 y el denominador puede ser considerado como una medida de la magnitud de los dos datos unitarios. Aunque Lance y Williams no insistieron en ello, parecería juicioso utilizar esta medida únicamente con valores no negativos, puesto que de otra manera se obtendrían distancias negativas.

Hay que tener en cuenta que como para cada para de datos unitarios o elementos, se utiliza distinto denominador, no siempre satisfará la desigualdad triangular y, entonces, no sería una metrica.

Una alternativa solución a esta métrica, la dio GOWER, que cambió el denominador y obtuvo.

$$8.\ d(X_j,X_k) = \sum_{i=1}^{n} \frac{|x_{ij} - x_{ik}|}{|x_{ij}| + |x_{ik}|}$$

III. DISTANCIA ABSOLUTA:

$$9.\ d(X_j,X_k) = \sum_{i=1}^{n} |x_{ij} - x_{ik}|$$

Es debida a CARMICHAEL y SNEATH quienes la justificaron alegando que cuando, por ejemplo, dos UTOS están definidos mediante dos variables cuyas unidades de escala tienen el mismo valor, dichos OTUS tendrán entre sí la misma distancia tanto si se consideran como unidades diferentes para cada variable, como si se les considera como una unidad aparte para una variable y tres unidades independientes sobre la otra.

IV. DISTANCIA DE MAHALANOBIS:

Esta distancia, que lleva el nombre de su autor quien la propuso en 1936, fue utilizada en técnicas de conglomeración por FRIEDMAN y RUBIN (1967) y por MCRAE (1971).

Una de sus expresiones es

1º) $d^2(X_j,X_k) = (X_j - X_k)^T W^{-1} (X_j - X_k)$ siendo W^{-1} la matriz inversa de la matriz de dispersión interior a los grupos o matriz de varianza-covarianza.

Esta distancia tiene la ventaja sobre la euclidea y la absoluta de tomar en cuenta correlaciones entre las variables y compensa o corrige su efecto. Cuando dichas correlaciones son nulas es equivalente a la distancia euclidea utilizando variables estandarizadas. Se considera, frecuentemente, como una distancia generalizada y es invariante bajo transformaciones lineales no negativas.

Otra expresión de esta distancia es la siguiente:

$$d(\bar{X}_j, \bar{X}_k) = \left[(x_{j1} - x_{k1}), \ldots, (x_{jp} - x_{kp})\right] \underset{\sim}{R}^{-1} \cdot \left[(x_{j1} - x_{k1}), \ldots, (x_{jp} - x_{kp})\right]^T$$

en donde p es la matriz inversa de la matriz p x p de correlaciones entre caracteres.

V. DISTANCIA GENERAL CON PESOS:

La asignación de pesos [5*] a las variables puede hacerse o como fruto de una primera evaluación de la importancia de cada variable o por otras distintas razones que estén relacionadas con el problema clasificatorio.

$$d_{ij} = \left[\sum_{k=1}^{n} W_k \, (x_{ik} - x_{jk})^2\right]^{\frac{1}{2}}$$

donde W_k son los pesos de las variables.

VI. COEFICIENTE DE DIVERGENCIA:

$$C.D. = \left[\frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_{ij} - x_{ik}}{x_{ij} + x_{ik}}\right)^2\right]^{\frac{1}{2}}$$

e está relacionado con la métrica de Canberra y que se debe a LARK (1952); es una medida heurística de distancia, aunque no es erdaderamente métrica.

I. DISTANCIA DE JEFFREYS-MATUSITA:

$$M = \left[\sum_{k=1}^{n} \left(\sqrt{x_{ki}} - \sqrt{x_{kj}} \right)^2 \right]^{\frac{1}{2}}$$

Es una medida heurística de distancia y que no es verdaeramente métrica, aunque ha sido utilizada con frecuencia. Esta istancia fue, originalmente, definida como una distancia entre os funciones de densidad de probabilidad, aunque en la forma daa se puede utilizar como una medida de distancia entre un par de UTOS.

VIII. COEFICIENTE DE PARECIDO RACIAL:

Fue desarrollado por K. Pearson (1926)

Aunque se incluyó este coeficiente entre los angulares, también figura aquí porque se suele utilizar como coeficiente de disimilitud.

En este caso se miden los caracteres continuos y son expresados como medias.

$$C.R.L. = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{(\bar{X}_{ij} - \bar{X}_{ik})^2}{\frac{S_{ij}^2}{t_j} + \frac{S_{ik}^2}{t_k}} \right]^{\frac{1}{2}} - \frac{2}{n}$$

donde $\bar{X}_{ij}$ representa la media muestral del carácter i-ésimo para la UTO j, S_{ij}^2, la varianza del mismo y t_j, el tamaño de la muestra UTO_j.

En taxonomía numérica se puede trabajar con la fórmula:

$$C.R.L. = \left| \frac{1}{2n} \sum_{i=1}^{n} (x_{ij} - x_{jk})^2 \right|^{\frac{1}{2}} - \frac{2}{n}$$

empleando datos del tipo Q que han sido estandarizados por filas ya que no tenemos medias, sino únicamente, valores únicos (con una varianza igual a uno), representando las UTOS.

IX. COEFICIENTE DE ROGERS Y TANIMOTO (1960):

$$d(X_j, X_k) = - \log_2 S_{jk}$$

siendo S_{ik} un coeficiente de asociación entre las UTOS j y k y los rangos, variando entre 0 y 1.

Estas distancias definen un espacio SEMIMETRICO, mejor que uno métrico, es decir, que no verifican la condición "d > 0 para i ≠ j).

Para medir la heterogeneidad o desemejanza, se puede emplear la fórmula siguiente

$$1)\ d(X_j, X_k) = \sqrt{\sum_{j=1}^{m} \frac{1}{P_i} \sum_{h=1}^{P_i} \left(\frac{x_{jhk} - x_{khi}}{x_{jhk} + x_{khi}} \right)^2}$$

en donde:

m= número de grupos de indicadores

P_i= número de indicadores del grupo k

x_{jhk} y x_{khi} = valores de los indicadores h del grupo i en los UTOS j y k respectivamente.

La expresión anterior es invariante a cambios de escala, así como a inversiones, pero no a cambios de origen ni complementaciones.

Si el número de grupos coincidiese con el número de indicadores 0, lo que es lo mismo, si hubiese un solo indicador en cada grupo, o no se ponderase, se verificaría $P_1 \ldots P_m = 1$, $P = m$, y se obtiene la fórmula:

$$2)\ d(X_j, X_k) = \sqrt{\sum_{h=1}^{m} \left(\frac{x_{jhk} - x_{khi}}{x_{jhk} + x_{khi}} \right)^2}$$

La medida de la desemejanza varía entre 0 y 1 para cada indicador positivo.

Con esta fórmula, la diferencia absoluta entre las magnitudes entre las cuales se produce, sea mayor o menor.

Las anteriores expresiones 1) y 2) tienen las siguientes ventajas:

a. TIPIFICACION: En general, la estandarización de los caracteres o indicadores es necesaria para que las medidas sean comparables, ya que los caracteres pueden expresarse en valores absolutos con unidades que pueden ser muy distintas en porcentajes, etc.

b. PONDERACION: La influencia de la magnitud de los indicadores está compensada, ya que cada diferencia se divide por la suma.

Además, en 1) se trata de compensar la variación en el número de indicadores de cada grupo, dividiendo por este número, P_k, el cuadrado de la diferencia dividida por la suma de los valores del indicador correspondiente en ambos países.

X. IVANOVIC:

Una expresión que tiene en cuenta las correlaciones entre caracteres y que se debe a IVANOVIC (1965) es la siguiente:

$$\sum_{h=1}^{P} \frac{|d_h|}{S_h} \quad \sum_{i=1}^{P-1} (1-r_{hi}),\ d_h = X_{jh} - X_{kh}$$

o también,

$$\sum_{h=1}^{P} \frac{|d_h|}{S_h} \quad \sum_{i=1}^{h-1} (1-r_{hi,\ 1,2,\ldots,i-1})$$

En donde S_h representa la desviación estadard del carácter o indicador h y r_{hi} la correlación entre los caracteres h,i.

Si se aplica la fórmula o corrección de IVANOVIC a la expresión, ya estudiada,

$$d(X_j, X_k) = \sqrt{\sum_{h=1}^{m} \left(\frac{x_{jhk} - \bar{x}_{khi}}{x_{jhk} + S_{khi}} \right)^2}$$

se obtiene:

$$d(X_j,X_k) = \sqrt{\sum_{h=1}^{P}\left(\frac{x_{jh} - x_{kh}}{x_{jh} + x_{kh}}\right)^2 \quad \sum_{i=1}^{P-1}(1 - r_{hi})}$$

XI. DIFERENCIAS EN TAMAÑO Y EN FORMA:

En el cálculo de medidas de similitud entre UTOS, PENROSE (1954) sugiere que se incluya el estudio de las diferencias en tamaño y en forma.

Penrose define el tamaño de una UTO como la media de sus valores de carácter y la distancia entre tamaños de dos UTOS, como el cuadrado de la diferencia de sus tamaños. Cuando esta distancia se resta de la llamada distancia MSD,[7*] esto es,

$$d^2(X_j,X_k) = \frac{1}{n}\left(\sum_{i=1}^{n} X_{ij}^2 + \sum_{i=1}^{n} X_{ik}^2 - 2\sum_{i=1}^{n} X_{ij}\,X_{ik}\right)$$

permanece un término que es la varianza de las diferencias en valores de caracter de los dos UTOS que han sido comparados.

PENROSE sugiere que esta varianza, que denominó DISTANCIA ENTRE FORMAS, se utilice como una medida de la diferencia de formas entre las dos UTOS, ignorando las diferencias de tamaño. Sin embargo, como indicaron ROHLF y SOKAL (1965), la distancia entre formas, únicamente ignora las diferencias aditivas de tamaño.

Las fórmulas para estas distancias de PENROSE son:

DIFERENCIA ENTRE TAMAÑOS:

$$Q_{jk}^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n} X_{ij} - \sum_{i=1}^{n} X_{ik}\right)^{2n}$$

DIFERENCIA ENTRE FORMAS:

$$z^2_{jk} = \frac{n}{n-1}\,(d^2_{jk} - Q^2_{jk})$$

en donde d^2_{jk} es la MSD comentada más arriba.

XII. COEFICIENTES DE DISIMILITUD QUE MIDEN LA VARIACION DENTRO DE LOS OTUS:

Existen varios coeficientes de semejanza que miden la variación dentro de los OTUS así como la diferencia entre las medias.

Similar al C.R.L. de Pearson es el coeficiente de SANGHUI:

$$T^2(X_j, X_k) = \frac{1}{n}\;\frac{\sum_{i=1}^{n} (\bar{X}_{ij} - \bar{X}_{ik})^2}{S^2_{ij} + S^2_{ik}}$$

Otro coeficiente es el de CROUELLO:

$$C.S.D. = \sum_{i=1}^{n} \left[(\bar{X}_{ij} - \bar{X}_{ik})^2 + (S_{ij} - S_{ik})^2\right]^{\frac{1}{2}}$$

en ambas fórmulas $\bar{X}_{ij}$ es la media y S_{ij} la desviación estándar del caracter i en el OTU j.

Para cuantificar la similitud entre muestras de poblaciones caracterizadas por frecuencias de genes EDWARDS y CAVALLI-SFORZA dieron la siguiente medida:

$$d(X_j, X_k) = \left| \sum_{i=1}^{m} (2 - 2 \cos \alpha_i) \right|^{\frac{1}{2}}$$

donde existe n lugares y, cuando para cualquier i, con m aleles y frecuencia P_g $\cos \alpha_i = \sum_{g=1}^{m} (P_{gj}, P_{gk})^{\frac{1}{2}}$.

Otro coeficiente, relacionado con el mismo tema fue propuesto por STEWART:

$$S = \frac{1}{2} \sum_{i=1}^{n} \left(\frac{\sum_{g=1}^{m} P_{gij} P_{gik}}{\sum_{g=1}^{m} P_{gij}^2 \sum P_{gik}^2 \; \frac{1}{2}} \right)^2$$

siendo n el número de locis y P_{gij} la frecuencia del alele g en el lugar i para el OTU j.

ROGERS desarrolló el siguiente coeficiente de distancia:

$$D = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{2} \sum_{g=1}^{m} (P_{gij} - P_{gik})^2 \right]^{\frac{1}{2}}$$

XIII. DISTANCIA DE CALHOUN:

BARTELS y otros propusieron una medida de distancia que utiliza únicamente la ordenación de los OTUS a lo largo de una ordenación de los OTUS a lo largo de una de las dimensiones del espacio.

El concepto básico para medir la distancia entre dos OTUS es imaginarlos como vértices opuestos de un hiperparalelepipedo cuyos lados son paralelos a los ejes del espacio. La distancia es básicamente la fracción de OTUS que "caen" o están dentro del hiperparalelepipedo y sus extensiones.

Sean: N_i el número de puntos que están situados entre los dos puntos de interés en, al menos, una variable, es decir, puntos del interior del hiperparalelepipedo y sus extensiones.

No es el número de puntos que no están entre los dos puntos de interés en cualquier dimensión, pero que es igual en valor, sobre una o más variables, con uno u otro de los OTUS.

N_z es el número de puntos que tienen valor igual sobre, al menos, una variable con los puntos de interés pero que no pertenecen ni al interior ni a la frontera del hiperparalelepípedo, es decir, los dos OTUS de interés son iguales sobre una o más variables de modo que la porción asociada del hiperparalelepipedo extendido es de espesor cero; entonces la distancia de CALHOUN es,

$$D_c = 6\ N_i + 3\ N_b + 2\ N_z$$

si el número total de OTUS es n, entonces el mayor valor posible de D_c es $6(n-2)$, en cuyo caso la distancia normalizada D_c es

$$D_c = (6\ N_i + 3\ N_b + 2\ N_z)/6(N-2)$$

La distancia de Calhoun no es una métrica, no cumple ni la desigualdad triangular ni el axioma 2; es invariante a transformaciones que conserven el orden de los OTUS a lo largo de los ejes de medida y no es invariante a las rotaciones ortogonales

MEDIDAS PROBABILISTICAS

Este tipo de medidas son de utilidad cuando sea conveniente utilizar la población estadística para modificar los pesos de las variables.

Cuando el peso está determinado por la población estadística, estas medidas muestran especial sensibilidad para la elección de la población. El peso depende del valor de la variable, al igual que la variable particular utilizada en la distancia afectada por pesos; por ejemplo, el peso o importancia del significado de la similitud será, en general, diferente si los objetos comparados toman los valores 1 y 1, cómo si son opuestos a tomar los valores 0 y 0. Mientras que los autores antiguos utilizaron los pesos como funciones de la población estadística, hoy en día, el usuario puede desarrollar su propio proceso para la asignación de los pesos.

La fórmula más importante para las medidas probabilísticas es,

$$\sum_{j=1}^{v} (\pm 1) \log \frac{1}{\text{prob.j } (x_j, y_j)}$$

siendo v el número de variables y tomando +1 si $x_j = y_j$ y -1 si $x_j \neq y_j$.

Dada una distribución de probabilidad para la ocurrencia de varios estados de caracter sobre n OTUS, se puede construir una medida de INFORMACION o ENTROPIA, aunque la información, en el sentido técnico no es una medida que nos ilustre o nos ayude a conocer mucho un grupo, aunque es una medida del desorden o confusión .

La medida del desorden para el carácter i es,

$$H(i) = - \sum_{g=1}^{m_i} P_{ig} \ln P_{ig}$$

donde m_i es el número de diferentes estados para el carácter i y P_{ig} es la proporción observada de los n OTUS que tienen o exhiben el estado g para el carácter i.

Si los n caracteres no están correlacionados, podemos sumar los valores separados de H(i) para conseguir la información total del grupo:

$$I = + \sum_{i=1}^{n} H(i)$$

Un índice basado sobre información mutua, ESTABROOK y ORLOCI, es el cociente de las sumas de información contenida exclusivamente en los caracteres h e i, y la información total que poseen h e i conjuntamente.

Bajo especiales condiciones todos estos coeficientes de desorden podrían ser métricos y puestos en la forma de coeficientes de similitud (coeficiente de coherencia de RAJSKI) de la siguiente forma:

$$S(h,i) = \left[1 - d^2(h,i)\right]^{\frac{1}{2}}$$

que será una medida de la similitud entre los caracteres h e i basada en una medida del desorden entre ellos.

Finalmente, diremos que el coeficiente de Rogers y Tanimoto, que hemos incluido en la sección anterior, es, en esencia, un coeficiente de similitud probabilístico y de información, su formulación fue $d(X_j,X_k) = - \log_2 S(X_j,X_k)$, donde d representaba la distancia entre dos OTUS j y k y S era un coeficiente de similitud entre los OTUS j y k; a partir de esta distancia y similitud definieron,

$$H_j = \sum_{k=1}^{t} d(X_j,X_k) = \sum_{k=1}^{t} - \log_2 S(X_j,X_k)$$

donde $k \neq j$, y S 0, que es una media del contenido de información (en bits) del OTU j.

Una medida de divergencia entre dos distribuciones de probabilidad P_i (x) y P_j(x) es el RADIO DE INFORMACION de SIBSON.

$$d_{ij} = \tfrac{1}{2}\int_x \left\{P_i(x) \log_2 \left[P_i(x)/q(x)\right] + P_j(x) \log_2 \left[P_j(x)/q(x)\right]\right\} d_x$$

siendo $q(x) = \tfrac{1}{2}\left[P_i(x) + P_j(x)\right]$

Otro tipo de medida, para variables binarias, puede derivarse de la tabla de concordancia-discordancia de un argumento probabilístico:

Si $\frac{A}{A_1+C}$ es la probabilidad condicional de que el OTU 2 tome el valor 1 sobre la variable elegida aleatoriamente, dado que el OTU 1 toma el valor 1 sobre esa variable y, análogamente la probabilidad $\frac{A}{A+B}$, entonces una medida simétrica de la similitud entre los dos objetos está dada mediante la fórmula,

$$d(\bar{X},\bar{Y}) = \tfrac{1}{2}\left[\frac{A}{A+C} + \frac{A}{A+B}\right]$$

MEDIDAS FUNCIONALES DE SIMILITUD

Estas medidas son funciones de las diferentes distancias y pueden expresarse mediante la siguiente fórmula general

$$S_{ij} = \frac{1}{1 + d_{ij}}$$

siendo d_{ij} una distancia.

En un estudio sobre contornos de isosimilitud elaborado por GEOFFREY H. BALL (12) nos demuestra que si una medida funcional de similitud es una función monótona decreciente de distancia, entonces la forma del contorno de isosimilitud es la misma que la de la distancia de la que la medida es función.

En general podrían definirse otras funciones monótonas decrecientes de las distancias, como medidas de similitud.

OTROS COEFICIENTES DE DISIMILITUD

Como complemento de este capítulo, en el que presentamos un catálogo de coeficientes de disimilitud, y ofrecemos, a continuación, algunas nuevas distancias y coeficientes de disimilitud que pueden ser base de futuros trabajos en este campo tan sugestivo de la taxonomía numérica.

A. Consideremos los elementos atmosféricos causantes de la polución en una determinada zona.

Si hacemos un tratamiento de dicha zona con determinados productos, podemos conseguir que precipiten esos elementos nocivos.

La cantidad precipitada cuando entran en contacto dos elementos x e y de distinto peso atómico, es proporcional a la diferencia de sus pesos, por lo que podemos definir una distancia d, entre el conjunto de los elementos y el cuerpo de los números reales de la siguiente forma:

$$d: E \times E \longrightarrow \mathbb{R}$$

$$d(x,y) = K|P_x - P_y|, \quad K > 0$$

Efectivamente, d es una distancia o métrica puesto que se verifican las siguientes propiedades:

1. $d(x,x) = K\,|P_x - P_x| = 0$

2. $d(x,y) = K\,|P_x - P_y| = K\,|P_y - P_x| = d(y,x)$

3. $d(x,y) + d(y,z) \_ d(x,z)$ ya que

$$K\left|P_x - P_y\right| + K\left|P_y - P_z\right| = K\left[\left|P_x - P_y\right| + \left|P_y - P_z\right|\right] \geq$$
$$\geq K\left[(P_x - P_y) + (P_y - P_z)\right] = K\left|P_x - P_z\right| = d\,(x,z)$$

La constante de proporcionalidad, K, podría servir para indicar la afinidad que existe entre los conjuntos formados por cada tipo de elementos nocivos.

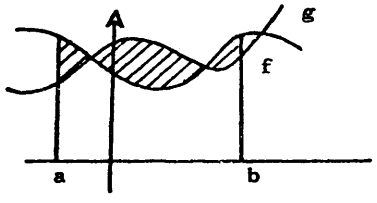
B. Sea C(a,b) la clase de funciones continuas definidas sobre el intervalo (a,b). Definimos las distancia d así:

$$d: C(a,b) \longrightarrow R$$

$$d(f,g) = \int_a^b \left|f(x) - g(x)\right| dx$$

Efectivamente, d es una métrica sobre C(a,b) ya que,

1. Si $f \neq g$, $d(f,g) = \int_a^b \left|f(x) - g(x)\right| dx \geq 0$



Precisamente, la integral anterior representa el área de la región rayada en la figura.

Si $f = g$ en $(a,b) \Leftrightarrow f(x) = g(x)$, $\forall\, x \in (a,b)$, y $f(x) - g(x) = 0$. luego, $d(f,g) = 0$. El área es cero

2. $d(f,g) = \int_a^b |f(x) - g(x)|\ dx = \int_a^b g(x) - f(x)\ dx =$
$= d(g,f)$

3. $d(f,h) \leq d(f,g) + d(g,h)$, ya que
$|f(x) - h(x)| = |f(x) - g(x) + g(x) - h(x)| \leq$
$\leq |f(x) - g(x)| + |g(x) - h(x)|$ y resulta que
$d(f,h) = \int_a^b |f(x) - h(x)|\ dx \leq \int_a^b |f(x) - g(x)|\ dx +$
$+ \int_a^b |g(x) - h(x)|\ dx = d(f,g) + d(g,h)$

La continuidad de las funciones en (a,b) nos asegura la existencia de las integrales, aunque, en realidad, solo hace falta que las funciones sean continuas en todos los puntos de (a,b) salvo en un número finito de punto de (a,b). En este supuesto siguen existiendo las integrales.

Dadas las funciones,

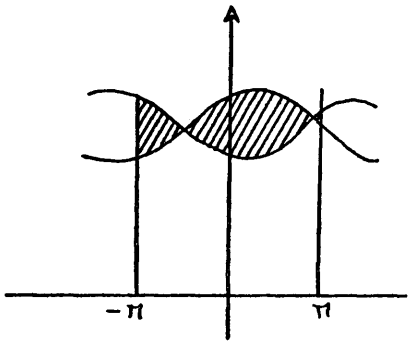
$$f(t) = \frac{x_1}{r_1} + x_2 \text{ sen } t + x_3 \cos t + \ldots$$

$$g(t) = \frac{x'_1}{r_1} + x'_2 \text{ sen } t + x'_3 \cos t + \ldots$$

definidas en $(-\pi, \pi)$, hemos de asegurarnos de que sean continuas, menos en un número finito de puntos, en $(-\pi, \pi)$ para que d, definida por:

$d(f,g) = \int_{-\pi}^{\pi} |f(t) - g(t)|\ dt$, sea efectivamente una distancia.

Gráficamente, d(f,g) es el número que expresa el área rayada de la figura:



Este número nos puede dar una idea de la aproximación de las funciones f y g en $(-\pi, \pi)$.

C. GEODESICAS

a. Consideraciones geométricas: Las Geodésicas se suelen definir como "las curvas de mínima distancia entre puntos de una superficie".

En el plano, las Geodésicas son rectas, que dan la mínima distancia entre dos cualesquiera de sus puntos.

En la superficie esférica "la distancia más corta entre dos puntos A y B se mide sobre una geodésica, que es un círculo máximo".

Ahora bien, hay dos arcos de círculo máximo que unen dos puntos, siendo solo uno de ellos la curva de mínima distancia, exceptuando el caso en que A y B sean extremos de un diámetro, ya qu entonces existen infinitos arcos, de la misma longitud. Es decir, no siempre es cierto que por dos puntos pase una sola geodésica.

En todo caso, para hallar la distancia más corta entre

A y B, si existe, hay que determinar la función,

$$v = v(u)$$

para que S tome el valor mínimo,

$$(dS^2 = E\,du^2 + 2F\,du\,dv + G\,dv)$$

Este es un problema clásico del cálculo variacional.

"Si existe un sólo arco de geodésica que une dos puntos, de una misma región, este arco da el camino más corto en esa región, que une estos puntos".

La cuestión de si una geodésica es la mínima distancia entre dos puntos fue planteada, por primera vez, por JACOBI.

b. Consideraciones taxonómicas: El análisis morfométrico es una nueva técnica que estudia viejos problemas: representa un acercamiento cuantitativo a los problemas taxonómicos.

Ha sido tal su desarrollo que han surgido multitud de métodos, cada uno de los cuales enfatiza en un determinado aspecto, y, por tanto, se hace necesaria su selección.

Entre dichos métodos están, naturalmente, los que factorizan la matriz de correlaciones entre taxones o grupos con respecto a sus caracteres (técnicas Q) y los que factorizan las correlaciones entre los caracteres con respecto a los taxones (técnicas R).

En este estado de cosas es muy importante considerar, como factor de peso, la generalidad de los mismos y, por tanto, surgen inmediatamente las distancias generalizadas y los métodos que permiten su cálculo, métodos que aplican la proposición según la cual "si estamos trabajando con variables correlacionadas en un espacio rectangular, podemos cambiar a un hiperespacio curvado con variables no correlacionadas".

Una aplicación de la mencionada proposición es la teoría general de la relatividad, con la que está estrechamente relacionada la técnica de la distancia generalizada (125).

La longitud de una geodésica en la teoría general está dada por la expresión,

$$dS^2 = \sum g_{ij}\, dx_i\, dx_j$$

siendo dx_i y dx_j las diferencias sobre los ejes x_i y x_j, respectivamente y g_{ij} el tensor fundamental que describe la curvatura del hiperespacio en el que se mide la geodésica.

La expresión análoga para la distancia generalizada entre dos grupos de organismos es

$$D^2 = \left|di\right|\, A^{-1}\, \left|dj\right| \qquad \text{o bien,}$$

$$D^2 = A^{-1} \left|di\right| \left|dj\right|$$

donde A es la matriz de correlaciones entre los caracteres y di_{ij} representa el vector de las diferencias medias de los caracteres para los dos grupos.

Por tanto, la matriz inversa de la matriz de dispersión o de correlación juega, esencialmente, el mismo papel que el tensor fundamental g_{ij} que describe la distorsión del hiperespacio de Riemann, al acomodar las relaciones que existen entre los caracteres cuando se miden en el espacio euclideo.

Esta relación se entiende cuando se sustituye el tensor fundamental g_{ij} por la delta de Kronecker, δ_{ij}, que es una matriz unidad en la que los términos de la diagonal principal representan varianzas estandarizadas, y los restantes elementos nulos, lo que indica la ausencia de correlación entre los caracteres.

La expresión de la distancia generalizada, de esta forma, se convierte en famoso resultado del teorema de Pitágoras:

$$D^2 = d_i^2 + d_j^2 + \ldots$$

en un espacio euclideo del mismo número de dimensiones que de caracteres.

Esta última expresión es una medida de distancia utilizada en taxonomía, pero que se suele utilizar cuando los caracteres no varían mucho de un individuo a otro dentro de un grupo.

En este caso, las covarianzas y las correlaciones no se hacen cada vez más pequeñas, como las varianzas, sino que se hacen indeterminadas.

E. Sea $X = x_1, x_2, \ldots, x_n$ un conjunto de n OTUS, cada uno de los cuales está definido por los p caracteres:

$$x_1 \longrightarrow x_{11}, x_{12}, \cdots, x_{1p}$$
$$x_2 \longrightarrow x_{21}, x_{22}, \cdots, x_{2p}$$
$$\vdots$$
$$x_n \longrightarrow x_{n1}, x_{n2}, \cdots, x_{np}$$

podemos considerar la OTU ficticia x cuyos caracteres son,

$$x \longrightarrow y_1, y_2 \cdots \cdots y_p \quad \text{siendo}$$

$$y_j = \min \left\{ x_{ij} / \; i = 1,2,\ldots, n \right\}_j \;, \; j = 1,2,\ldots,p$$

y hallar las distancias máximas o mínimas entre los elementos del conjunto X y la OTU ficticia x y, de esta forma, podemos ordenar jerárquicamente, según dichas distancias, el conjunto de elementos X con objeto de tomar alguna decisión sobre los elementos de X.

Si X = países y x_i representan índices económicos, x sería el país menos desarrollado y así podríamos ordenar los países según su riqueza o potencial económico.

Si consideramos como x el elemento ficticio cuyos caracteres construimos así:

$$y_j = \frac{\max \left\{ x_{ij} / i = 1,2,\ldots,n \right\} + \min \left\{ x_{ij} / i = 1,2,\ldots,n \right\}}{2}$$

donde j= 1,2,... ... p, construiríamos un elemento medio ideal que también nos permitiría obtener una ordenación de los elementos u OTUS, a ambos lados de x, que también nos serviría de pauta para tomar distintas decisiones sobre unos y otros elementos.

F. Es evidente que lo perfecto no existe, que es imposible encontrar la mejor distancia o el mejor coeficiente de disimilitud, pero el científico ha de investigar procurando acercarse hacia ese ideal, y una de las formas en que se suele llevar a cabo esa búsqueda es a base de variaciones sobre coeficientes ya establecidos; por ejemplo si $d(\bar{X},\bar{Y}) = \sqrt{\Sigma(x_i-y_i)^2}$ es la distancia euclidea entre los elementos u OTUS X e Y, cuyos caracteres son, respectivamente, x_i e y_i, i = 1 ... ... p, podríamos tratar de considerar algunas variaciones sobre d como las siguientes:

$$i_1\ (X,Y) = \frac{\sqrt{\Sigma(x_i,y_i)^2}}{x_i}$$

$$i_2\ (X,Y) = \frac{\sqrt{\Sigma(x_i-y_i)^2}}{\left(\frac{\Sigma x_i+\Sigma y_i}{2}\right)} = 2\,\frac{\sqrt{\Sigma(x_i-y_i)^2}}{\Sigma x_i + \Sigma y_i}$$

$$i_3\ (X,Y) = \frac{\sqrt{\Sigma(x_i-y_i)^2}}{\sqrt{\Sigma(x_i+y_i)^2}}$$

$$i_4\ (X,Y) = \frac{\sqrt{\Sigma(x_i-y_i)^2}}{\sqrt{\Sigma x_i^2} + \sqrt{y_i^2}}$$

Los anteriores coeficientes de disimilitud, entre otros, nos permitirán estudiar propiedades similares a las ya establecidas para las métricas y es posible que en algún problema concreto, considerániolas como disimilitudes entre OTUS, siendo el soporte de algún nuevo método de clasificación, produjeran mejores resultados que los originados por los coeficientes clásicos.

1* : OPERATIONAL TAXONOMIC UNIT.

2* : Desde ahora, consideraremos como sinónimos, variables y caracteres.

3* : También se la llama MATRIZ TAXONOMICA BASICA.

4* : Como se advierte más adelante, se pueden tomar las variables estandarizadas de modo que sean independientes de las unidades.

5* : Hay autores que no son partidarios de atribuir pesos, siguiendo la doctrina de M. Adanson (véase cap. I).

6* : Definición de E. PFLAUMAN y H. UNGER (96).
Según KELLEY en "TOPOLOGIA GENERAL", EUDEBA, 1962, una pseudométrica no puede tomar valores infinitos.
En la bibliografía francesa suele denominarse "écart" a una función con las propiedades de una métrica, pero que puede tomar valores infinitos.

7* : MEAN SQUARE DISTANCE.

CAPITULO III

RELACIONES ENTRE CONGLOMERADOS

DISTANCIAS ENTRE CONJUNTOS

DISTANCIAS ENTRE CONGLOMERADOS

I. METODOS AGLOMERATIVOS:

II. METODOS DIVISIVOS:

SIMILITUD ENTRE CONGLOMERADOS

DISTANCIAS ENTRE DISTRIBUCIONES

DIVERGENCIA ENTRE DISTRIBUCIONES

HOMOGENEIDAD DE CONGLOMERADOS

FUNCION DE COHESION

DIFUSION Y CONEXION

AGRUPACIONES DE MAXIMA HOMOGENEIDAD

I. EL PROBLEMA SIN RESTRICCIONES:

II. EL PROBLEMA CON RESTRICCIONES:

LA DISTANCIA δ ENTRE CONGLOMERADOS

Al igual que entre individuos, objetos, entidades o - OTUS se habla de SIMILITUD o de desemejanza, parece natural estudiar estos conceptos entre conglomerados o clases.

El problema de la conglomeración exige, al menos, dos medidas, una de homogeneidad dentro de un conglomerado y otra de heterogeneidad entre conglomerados. Estas medidas están relacionadas con la noción de distancia entre conglomerados.

Ofrecemos, a continuación, una serie de procedimientos para medir distancias entre conglomerados, distancias entre distribuciones y distancias para conseguir agrupaciones de máxima - homogeneidad y, finalmente, los estudios que se han hecho con respecto a los conceptos de COHESION, DIFUSION y CONEXION que tan relacionados están con éstos procedimientos.

En primer lugar exponemos algunas ideas relacionadas - con el concepto de distancias entre conjuntos.

DISTANCIAS ENTRE CONJUNTOS

La noción de distancia entre puntos, que hemos estudiado en el capítulo II, se relaciona con la de longitud, en particular con la de longitud de intervalos que, inmediatamente,se extiende a a la noción general de medida. Así, por ejemplo, definida la medida como "una función de conjunto tal que la medida de la unión de conjuntos disjuntos es igual a la suma de las medidas de dichos conjuntos", se tiene la aplicación inmediata de este concepto a la distancia euclidea (longitud) de intervalos.

Se ha definido la distancia entre conjuntos como el extremo inferior de las distancias entre todos los pares posibles que pueden formarse con un elemento del primer conjunto y uno del segundo.

También se han manejado otros conceptos que pueden, o no, ser coincidentes con el anterior, como, por ejemplo, el siguiente:

> "Si X_i y X_j son dos subconjuntos de un conjunto X, se llama DISTANCIA, D_{ij}, entre X_i y X_j la medida de su diferencia simétrica, es decir:

$$D_{ij} = f\left\{(X_i - X_j) \cup (X_j - X_i)\right\} = f\left\{(X_i \cup X_j) \cap (\overline{X_j \cup X_i})\right\}"$$

Es evidente que D_{ij}, que es una métrica, satisface las condiciones de una función de medida entre conjuntos y que exponemos a continuación:

"Sea $S = \left\{X_1, X_2, \ldots \ldots, X_n\right\}$ un conjunto de conjuntos

numerables. Una función f se dice que es la función de medida del conjunto X si se cumplen las siguientes condiciones:

1. $f(X_i)=0, \quad \forall X_i \in X$

2. $f(\emptyset)=0$

3. $\begin{cases} \text{si } X_i \cap X_j=\emptyset, \ f(X_i \cup X_j)=f(X_i)+f(X_j) \\ \text{si } X_i \cap X_j \neq \emptyset, \ f(X_i \cup X_j)=f(X_i)+f(X_j)-f(X_i \cap X_j)" \end{cases}$

Este concepto de función de medida satisface nuestra idea intuitiva de medida de un conjunto al que pertenecen un número finito de elementos, identificando $f(X_i)$ con el número de elementos del conjunto X_i.

Como ejemplo de la aplicación de las ideas anteriores a diversas ciencias citaremos el siguiente principio que utiliza el término SEMEJANZA como base para clasificar objetos en Psicología:

"Dos objetos están psicológicamente CERRADOS (o SEPARADOS) cuando son SEMEJANTES (o cuando no lo son".
Para establecer el concepto de que "dos elementos están psicologicamente cerrados o separados" hemos de basarnos en la distancia entre dos conjuntos X_i y X_j es decir en "la medida del conjunto constituido por los elementos no comunes a X_i y X_j, o sea, "la medida del conjunto $X_i \Delta X_j = (X_i-X_j) \cup (X_j-X_i)$".

DISTANCIAS ENTRE CONGLOMERADOS

I. METODOS AGLOMERATIVOS:

Sean $A = \{X_i / \ i = 1,\ldots,n\}$ y $B = \{Y_j / \ j = 1,\ldots,m\}$ dos conglomerados cuyos elementos pertenecen a una población P.

Sea $D = \{ d(X_i, Y_j) \}$, $i = 1,\ldots,n$; $j = 1,\ldots,m$, el conjunto de las distancias, entonces tenemos:

1. $D_1(A,B) = \min \{d(X_i, Y_j)\}$, $i = 1,\ldots,n$, $j = 1,\ldots,m$

   "La distancia entre dos conglomerados es la distancia entre los puntos más próximos de ambos conglomerados".
   Esta definición coincide con la definición de distancia entre conjuntos finitos.
   Esta medida es la base del método UNION SENCILLA, DISTANCIA AL VECINO MAS PROXIMO, ENLACE SIMPLE o ENLACE DEBIL.

2. $D_2(A,B) = \max \{d(X_i, Y_j)\}$ $i = 1,\ldots,n$, $j = 1,\ldots, m$

   "La distancia entre dos conglomerados es la distancia entre los dos puntos más alejados de dichos conglomerados".
   Esta medida es la base del método UNION COMPLETA, DISTANCIA AL VECINO MAS ALEJADO, ENLACE COMPLETO o ENLACE FUERTE.

3. $D_3(A,B) = \sum_{j=1}^{m} \sum_{i=1}^{n} d(X_i,Y_j)/n.m$ es la distancia media entre A y B con respecto a la función de distancia.

"La distancia entre dos conglomerados, es la distancia media entre todos los pares de individuos, tomando uno de un conglomerado y otro del otro conglomerado".

4. "La distancia entre dos conglomerados es la distancia entre las medias (o centroides, o centros de gravedad) de dichos conglomerados".

5.- "La distancia entre dos conglomerados es la distancia entre sus medianas o centros medianos".
Cada promedio (moda, media aritmética, etc.) dará origen a una definición de distancia.

6. Si $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ e $\bar{Y} = \sum_{i=1}^{m} \frac{Y_i}{m}$, y consideramos la MATRIZ DE DISPERSION:

$$\frac{n.m}{n+m} (\bar{X} - \bar{Y}) (\bar{X} - \bar{Y})^T$$

entonces la traza de esta matriz se llama DISTANCIA ESTADISTICA o SUMA DE CUADRADOS entre conglomerados A y B y se representa por:

$$D_y(A,B) = \frac{n.m}{n+m} (\bar{X} - \bar{Y})^T (\bar{X} - \bar{Y})$$

"La distancia entre dos conglomerados es la suma de

los cuadrados de las distancias de los elementos a la media del conglomerado total menos la suma de - los elementos a las medias de sus respectivos conglomerados".

Esta medida es la base del método de WARD.

LANCE y WILLIAMS (23) nos ofrecen la siguiente fórmula general para el cálculo de las DISIMILITUDES entre los grupos h y un grupo (j,k) formado por la fusión de los grupos j y k:

$$d_{(jk)} = \alpha_j \, d_{jh} + \alpha_k \, d_{kh} + \beta d_{jk} + \gamma d_{jh} - d_{kh}$$

figurando en la siguiente tabla, los valores de los parámetros - para distintas estrategias:

METODO	α_j	α_k	β	γ	AUTORES
ENLACE SENCILLO DISTANCIA MINIMA	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	SOKAL y SNEATH (1963)
ENLACE COMPLETO DISTANCIA MAXIMA	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	SNEATH y SOKAL (1963) MCQUITTY (1964)
DISTANCIA PROMEDIO NO PONDERADO	$\frac{n_j}{n_j + n_k}$	$\frac{n_k}{n_j + n_k}$	0	0	SOKAL y MICHENER (1958) MCQUITTY (1964)
DISTANCIA PROMEDIO PONDERADO	$\frac{1}{2}$	$\frac{1}{2}$	0	0	MCQUITTY (1966)
CENTROIDE	$\frac{n_j}{n_j + n_k}$	$\frac{n_k}{n_i + n_j}$	$\frac{-n_j \cdot n_k}{(n_j + n_k)^2}$	0	SOKAL y MICHENER (1958) GOWER (1967)
DISTANCIA MEDIANA	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0	GOWER (1967)
SUMA DE CUADRADOS	$\frac{n_j + n_k}{n_j + n_k + n_h}$	$\frac{n_k + n_l}{n_j + n_k + n_h}$	$\frac{-n_l}{n_j + n_k + n_h}$	0	WISHART (1969) ANDERSON (1971)

Si en lugar de conocer las disimilitudes, d_{ij}, conocemos las similitudes, S_{ij}, subsisten las mismas relaciones, sin más que tener en cuenta que,

$$S_{ij} = 1 - d_{ij}$$

y si $\alpha_i + \alpha_2 + \beta = 1$, se puede sustituir d por S directamente.

Las relaciones que se deducen de la fórmula de LANCE y WILLIAMS se aplican, generalmente, en el proceso de conglomeración como un procedimiento aglomerativo y, a partir de una matriz de similitud o disimilitud entre entidades u OTUS.

II. METODOS DIVISIVOS:

1. MONOTETICOS: Se estudian las variables hasta encontrar la dicotomía que maximiza algún criterio de disimilitud. Un criterio bastante utilizado es

$$\sum_{\substack{j=1 \\ j\neq k}}^{d} \chi^2_{jk} \quad , \quad k=1,\ldots \ d$$

siendo χ^2_{jk} el coeficiente de asociación ji-cuadrad entre las variables X_j y X_k calculado a partir de la tabla (2 x 2) de la distribución marginal. "Se sigue dividiendo el conglomerado hasta encontrar un valor de k que maximiza la expresión anterior".

2. POLITETICOS: El coeficiente de disimilitud de MACNAUGHTON-SMITH para variables binarias es:

"Sea X_{Aj} la proporción de objetos en un grupo A que tienen asignado el uno en la variable j y sea X_{Bj} la proporción de apareamiento para un grupo B. Entonces la disimilitud entre los grupos A y B es

$$\sum_{j} (X_{Aj} - X_{Bj})^2 \sum_{k \neq j} \chi^2_{jk}$$

donde los χ^2_{jk} se calculan para el grupo combinado A + B.

SIMILITUD ENTRE CONGLOMERADOS

La conglomeración estratificada jerárquica es la representación definida con más cuidado y fue dada independientemente por CONSTANTINESLU (1966), HARTIGAN (1967), JARDINE (1967), JOHNSON (1967) y LERMAN (1970), (23).

Formalmente, "es un árbol r = (E,A,T) que consiste en una raiz (origen) E, (el conglomerado que contiene todos los elementos, un conjunto finito, A, de nudos C (conglomerados de elementos) y una aplicación T de A en sí mismo, tal que para todo $K \geq 1$ $T^k C = C$ si y solo si C = E junto a una función de valores reales σ sobre A tal que $\sigma(C) \leq \sigma(C')$ si existe un $K \geq 0$ tal que $T^k C' = C$."

La representación pictórica de un dendrograma ha sido ampliamente utilizada para describir clasificaciones.

"La SIMILITUD, S_{ij}, entre dos conglomerados (o dos elementos) se define como el valor de σ en el primer nudo en el que dichos conglomerados quedan unidos por T."

"La DISIMILITUD, $d_{ij} = 1 - S_{ij}$, es una ultramétrica que satisface:

$$d_{ij} \leq \max \left[d_{ik}, d_{jk} \right] \quad \forall\, i,j,k. \text{ "}$$

DISTANCIAS ENTRE DISTRIBUCIONES

Consideremos cada conglomerado de puntos como una muestra extraída de una población.

Sean f y g las funciones de densidad de probabilidad correspondientes a los conglomerados A y B.

WACKER y LANGREBE discutieron diversas formas plurivalentes de medidas de distancia y sus propiedades métricas. Sus resultados figuran en la tabla que ofrecemos a continuación en la que C indica la clase de todas las funciones de distribución p-variantes absolutamente continuas, PN, la clase de funciones plurivariantes de distribución normal y PN_{Σ} la clase de funciones plurivariantes de distribución normal con matrices de igual covarianza.

La tabla también nos da las propiedades métricas de las medidas de distancia relativas a estas tres clases de funciones de distribución.

Algunas medidas de distancia entre conglomerados podrían resultar más útiles en aquellos casos en los que se puedan suponer normalidad y, entonces, se podrían estimar μ y Σ con $\bar{X}$ y S^2 y, de este modo, sería posible calcular dichas medidas.

Muchas de las medidas de la mencionada tabla subsisten para formas univariantes de las medidas de distancia (35).

FORMAS MULTIVARIANTES DE MEDIDAS DE DISTANCIA Y SUS PROPIEDADES METRICAS

1. CRAMER- VON MISES

$$w = \left\{ \int_{\infty}^{\infty} G(x) - F(x)^2 \, dx \right\}^{\frac{1}{2}}$$

2. KOLMOGOROV - SMIRNOV

$$K = \sup_{x} \left| G(x) - F(X) \right|$$

3. DIVERGENCIA

$$J = \int_{\infty}^{\infty} \lg_e \frac{f(x)}{g(x)} \left| f(x) - g(x) \right| dx$$

4. BHATTACHARYYA

$$B = \lg_e \int_{-\infty}^{\infty} \left| f(x) \, g(x) \right|^{\frac{1}{2}} dx$$

5. JEFFREYS - MATUSITA

$$M = \int_{-\infty}^{\infty} \left[\sqrt{g(x)} - \sqrt{f(x)} \right]^2 dx \quad ^{\frac{1}{2}}$$

6. DISTANCIA DE KOLMOGOROV

$$K_{(p)} = \int_{-\infty}^{\infty} \left| P_g \, g(x) - P_f \, f(x) \right| dx$$

7. NUMEROS DE KULLBACK - SIEBLER

$$L_{fg} = \int_{-\infty}^{\infty} \log_e \left(\frac{f(x)}{g(x)} \right) f(x) \; dx$$

8. SWAIN - FU

$$T = \frac{\mu_f - \mu_g}{D_f + D_g}$$

$$D_f = \left\{ \frac{\mu_f - \mu_g}{t_r \, \Sigma_f \, (\mu_f - \mu_g)(\mu_f - \mu_g)^T} \right\}^{\frac{1}{2}}$$

9. MAHALANOBIS

$$\Delta = \left\{ (\mu_g - \mu_f)^T \Sigma^{-1} (\mu_g - \mu_f) \right\}^{\frac{1}{2}}$$

10. SAMUELS - BACHI

$$\mu = \left\{ \int_0^1 \left[F^{-1}(\alpha) - G^{-1}(\alpha) \right] \; dx \right\}^{\frac{1}{2}} \quad \text{donde,}$$

$$F^{-1}(\alpha) = \inf \left\{ C / Q_c \cap Q \neq \emptyset \right\} \; y$$

$$Q_c = \left\{ x / \sum_{i=1}^{p} x_i \leq C \right\}_j \; , \; Q_\alpha = \left\{ x / F(x) \geq \alpha \right\}$$

11. KIEFER - WOLFOWITZ

$$V = \int_{-\infty}^{\infty} \left| F(x) - G(x) \right| \, e^{-|x|} \; dx$$

donde $|x|$ = vector norma de x

DIVERGENCIA ENTRE DISTRIBUCIONES

Una medida de la Divergencia entre dos distribuciones de probabilidad nos la ofrece SIBSON bajo el nombre de RADIO DE INFORMACION (54).

Sean las distribuciones $p_i(x)$ y $p_j(x)$,

$$d_{ij} = \frac{1}{2} \int_x \left\{ p_i(x) \log_2 \left| p_i(x)/q(x) \right| + p_j(x) \log_2 \left| p_j(x)/q(x) \right| \right\} dx$$

donde $q(x) = \frac{1}{2} \left| p_i(x) + p_j(x) \right|$

Esta medida puede ser sumada sobre diferentes variables X

A menudo, es conveniente categorizar variables numéricas, así como obtener distribuciones discretas. En este caso la integración en la anterior expresión se sustituye por el sumatorio extendido a los diferentes estados.

HOMOGENEIDAD DE CONGLOMERADOS

Sean (i) y (k) los dos conglomerados con menor distancia (en la estrategia utilizada) en el nivel k, es decir, los más similares, por tanto, en el nivel k+1, el número de grupos sería C-1, siendo C el número de grupos formados en el nivel k. Si al formar el nuevo grupo (ik) el incremento en la distancia total J, siendo,

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left\| x_j^{(i)} - m^{(i)} \right\|^2$$

es muy fuerte, se puede considerar que los grupos (i) y (k) no son HOMOGENEOS y que, por tanto, pertenecen a colectivos o a poblaciones, o a distribuciones probabilísticas distintas (38).

En cambio, si el incremento en J es pequeño, en principio, se puede admitir la "hipótesis nula" de que ambos grupos pertenecen al mismo colectivo y, por tanto, que la nueva agrupación (ik) es correcta. ¿Dónde está el límite en base al cual se decide el agrupamiento o, por el contrario, se estima que ambos grupos no son HOMOGNEOS, y al no serlo los demás, el dendrograma debe finalizar?

La respuesta a esta pregunta depende de la distribución probabilística seguida por la variable J(ik).

$$J_{(ik)} = \sum_{i=1}^{n_i} \left\| x_j^{(i)} - m^{(ik)} \right\|^2 + \sum_{k=1}^{n_k} \left\| x_j^{(k)} - m^{(ik)} \right\|^2$$

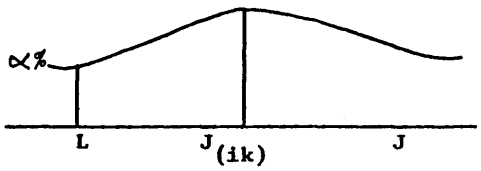
Antes de efectuar la fusión de los grupos (i) y (k), la suma total de las desviaciones euclideas de ambos grupos será:

$$J_{(ik)} = \sum_{j=1}^{n_i} \left\| x_J^{(i)} - m^{(i)} \right\|^2 + \sum_{j=1}^{n_k} \left\| x_j^{(k)} - m^{(k)} \right\|^2$$

La distancia $J_{(i+k)}$ será siempre menor o igual que la distancia $j_{(ik)}$.

Si los grupos (i) y (k) son HOMOGENEOS y, por tanto, pertenecen al mismo colectivo, la variable muestral $J_{(i+k)}$ es una buena estimación de la variable $J_{(ik)}$, y tiene el promerio $\bar{J}_{(ik)}$ cuya estimación es $J_{(ik)}$. Si $J_{(i+k)} \ll J_{(ik)}$ se puede suponer que la variable $J_{(i+k)}$ pertenece a un colectivo cuyo promedio es inferio a $\bar{J}_{(ik)}$ y que, por tanto, al no ser (i) y (k) grupos homogéneos, no pueden fusionarse.

Para obtener el límite L tal que para un α por 100 de riesgo se supone



que si $J_{(i+k)} < L$ los grupos no son homogéneos, se procede del siguiente modo:

1. Obtención de las distancias $J_{(i+k)}$ y $J_{(ik)}$.

2. Obtención del valor tipificado U en la ley normal para un α por 100 de riesgo.

3. La tipificación del límite L en la ley normal para el promedio $\bar{J}_{(ik)}$ y varianza $S_{J_{(ik)}}$ será:

$$\frac{\bar{J}_{(ik)} - L}{S_{J_{(ik)}}} = U_\alpha \qquad \frac{J_{(ik)} - L}{\sqrt{\frac{2}{(n_i + n_k)F} \cdot J_{(ik)}}} = U_\alpha$$

siendo,

$$\bar{J}_{(ik)} = (n_i + n_k).F.S$$

$$S_{J_{(ik)}} = \frac{2}{(n_i + n_k)F} \; J^2_{(ik)}$$

F = número de características.

$n = n_i + n_k$

S = promedio de la estimación de la estimación de las varianzas.

Por tanto, $L = J_{(ik)} - J_{(ik)} \cdot \sqrt{\frac{2}{(n_i + n_k)F}}$

1. El test de hipótesis consiste en analizar si

$$\frac{J_{(i+k)}}{J_{(ik)}} < 1 - U_\alpha \cdot \sqrt{\frac{2}{(n_i + n_k)F}}$$

de tal forma que si así ocurre, se supone con un por 100 de riesgo que los grupos son HETEROGENEOS y, por tanto, no se debe continuar en la obtención de nuevos niveles en el endograma.

Ahora bien, la prueba anterior tiene el inconveniente de que se basa en la hipótesis de que los elementos a agrupar siguen la ley multivariante normal.

A continuación, se ofrece un test de significación F de Fisher aplicable sea cual fuere la distribución probabilística de los elementos a agrupar.

Vamos a estimar si una distribución de elementos en C_2 grupos es significativamente mejor que una subdivisión en C_1 grupos donde $C_1 < C_2$. Para ello, se parte de la base de que se conocen los valores muestrales de las desviaciones euclideas $J(C_1)$ y $J(C_2)$,

$$J(C_1) = \sum_{i=1}^{C_1} \sum_{j=1}^{n_i} \left\| X_j^{(i)} - m^{(i)} \right\|^2$$

$$J(C_2) = \sum_{i=1}^{C_2} \sum_{j=1}^{n_i} \left\| X_j^{(i)} - m^{(i)} \right\|^2$$

Dado que $C_1 < C_2$, resulta que $J(C_1) \geq J(C_2)$. La variable muestral en el test de Fisher será:

$$\frac{J(C_1) - J(C_2)}{J(C_2)}$$

El proceso a seguir es el siguiente:

A. Obtención del valor teórico de la F (de Fisher), para $K_1 = F$ y $K_2 = F.(n-C)$ grados de libertad al nivel .

B. Obtención de J(C)

$$J(C) = \sum_{j=1}^{n_i} \left\| X_j^{(i)} - m^{(i)} \right\|^2 + \sum_{j=1}^{n_k} \left\| X_j^{(k)} - m^{(k)} \right\|^2 + \sum_{\substack{a=1 \\ a \neq i}}^{C} \sum_{\substack{j=1 \\ a \neq k}}^{n} \left\| X_j^{(a)} - m^{(a)} \right\|^2 .$$

C. Obtención de $m^{(ik)}$

$$M^{(ik)} = \frac{n_i \cdot m^{(i)} + n_k \cdot m^{(k)}}{n_i + n_k}$$

D. Obtención del valor real de F de Fisher,

$$Fr = \frac{\sum_{j=1}^{n_i} \left\| X_j^{(i)} - m^{(ik)} \right\|^2 + \sum_{j=1}^{n_k} \left\| X_j^{(k)} - m^{(ik)} \right\|^2 - \sum_{j=1}^{n_i} \left\| X_j^{(i)} - m^{(i)} \right\|^2 - \sum_{j=1}^{n_k} \left\| X_j^{(k)} - m^{(k)} \right\|^2}{\sum_{i=1}^{C} \sum_{j=1}^{n_i} \left\| X_j^{(i)} - m^{(i)} \right\|^2} \left[\left(\frac{C}{C-1} \right)^{2/F} - 1 \right]$$

E. Comparación de Fr y Ft

Si Fr Ft se puede asumir con un α por 100 de riesgo que los grupos considerados no son homogeneos y, por tanto, no deben fusionarse.

Si Fr / Ft no hay indicios estadístico como para considerar que J(C-1) es significativamente mayor que J(C); por tanto ,se fusionann los dos grupos considerados en el nivel K, alcanzando en el endograma el nivel K+1 formado por C-1 grupos.

FUNCION DE COHESION

En el estudio de los métodos de aglomeración (clumping) NEEDHAM definió una FUNCION DE COHESION, que mide la extensión a qué objetos del aglomerado (clump) están relacionados con objetos del complemento y buscó particiones que correspondienran a mínimos locales de esta función de cohesión simétrica entre un aglomerado y su complemento.

Si S_{ij} es la similitud entre los objetos i y j y si

$$S(A,B) = \sum_{i \in A} \sum_{j \in B} S_{ij}$$

mide la similitud entre los conglomerados A y B, entonces la función de cohesión entre un aglomerado C, que contiene n_c objetos y su complemento $\bar{C}$, es la siguiente:

$$\frac{S(C,\bar{C})}{S(C,C)} \left| \frac{n_c(n_c-1)}{S(C,C)} - \frac{S(C,C)}{\alpha n_c(n_c-1)} \right|$$

en donde α es un parámetro que controla el grado de solapamiento entre diversos aglomerados.

DIFUSION Y CONEXION

ESTABROOK ha propuesto formas de medir el aislamiento y a cohesión de los grupos que resultan de un estudio clasificatorio definió la DIFUSION y la CONEXION de un grupo (54).

"La DIFUSION de un grupo G es la diferencia entre el nivel h_1, nivel al que el grupo se formó y el nivel h_2, nivel al que el grupo se incorporó o unió a otro grupo mayor".

Si el grupo contiene N_G objetos, el número de segmentos debe estar comprendido entre un mínimo, N_G, y un máximo, $\frac{1}{2} N_G(N_G-1)$.

Si el número de segmentos es e, la conexión se define por:

$$C = \frac{e - (N_G - 1)}{\frac{1}{2}N_G(N_G-1)-(N_G-1)}$$

AGRUPACIONES DE MAXIMA HOMOGENEIDAD

W.D.FISHER (45) estudia, también, agrupaciones de máxima HOMOGENEIDAD y consigue un procedimiento práctico para agrupar elementos buscando la minimización de la varianza dentro de los grupos.

El problema estadístico, de gran importancia, como es el de definir grupos homogéneos, se plantea así:

"Dado un conjunto de K elementos, cada elemento tiene asignado un peso w_i y una medida numérica a_i, dado un entero positivo G, $G < K$, hay que hallar un procedimiento sistemático y práctico para agrupar los K elementos en G subconjuntos mutuamente excluyentes y exhaustivos, tales que la suma DE CUADRADOS PONDERADA,

$$D = \sum_{i=1}^{K} w_i (a_i - \bar{a}_i)^2$$

sea MINIMA, siendo $\bar{a}_i$ la media aritmética ponderada de aquellos a_i que son asignados al subconjunto al que se asignó el elemento i".

Este problema ha sido llamado "PROBLEMA DE AGRUPACION".

El valor D, que se llama "suma de cuadrados intergrupos" en el sentido del análisis de la varianza, será llamado DISTANCIA CUADRADA.

El problema anterior puede plantearse, desde un punto de vista geométrico, de la siguiente forma:

"Dados K puntos ponderados sobre una linea recta, agrupar

los puntos en G grupos de modo que la suma de distancias cuadradas de los puntos a sus centros de gravedad de grupo, sea mínima".

Un sistema de agrupación se llama una PARTICION, y una partición asociada con la distancia cuadrática mínima, D, se llamará "la menor partición cuadrática.

Se distinguen dos subclases en el problema de la agrupación:

1. El problema "SIN RESTRICCIONES", donde no se ponen restricciones o condiciones laterales a las particiones admitidas.

2. El problema "CON RESTRICCIONES", donde la concicionnes anteriores se imponen "a priori" sobre la base de un conocimiento previo, bien sea teórico, o bien por razones de conveniencia.

FISHER supone que w_i y a_i son conocidos y que la MEDIDA DE HOMOGENEIDAD, D, es útil para muchos problemas prácticos.

I. EL PROBLEMA SIN RESTRICCIONES:

Se puede plantear así: supongamos que se desea encontrar el mejor método para conseguir un número dado de estratos en un muestreo con estratificación proporcional, según una variable relevante de una determinada población, es decir, hay que restratificar una población, de modo que la estimación de la media de la población basada en una muestra estratificada tenga una varianza pequeña, suponiendo que se toma como media muestral, la estimación de la media de toda la población.

Ha sido demostrado que, bajo las condiciones anteriores, si w_i es el peso del elemento i de la población, a_i, la media de la clase i y $\bar{a}$ la media del estrato al que se asigna la clase i, entonces,la varianza de la estimación es proporcional a D, más una constante que representa la varianza dentro de las clases originales. Para minimizar la varianza de la estimación, es suficiente minimizar D que representa la varianza de las medias de los elementos dentro de los estratos finales.

Es obvio que cuando los K elementos originales se ordenan según la variable elegida (i j si $a_i < a_j$) las únicas particiones que se deben considerar son las CONTIGUAS definidas para un conjunto de elementos completamente ordenados según una partición que consiste en subconjuntos que satisfacen la siguiente condición: "si los elementos i,j y k están en el orden $i < j < k$ y los elementos i y k se asignan, al mismo estrato, también lo debe ser el elemento j. Para encontrar la agrupación óptima, es suficiente calcular los valores de D para cada una de $\binom{K-1}{G-1}$ particiones contiguas de los elementos en los K grupos elegidos y seleccionar la que tenga el menor valor de D.

RESUMEN: Elegida una variable de una población, al resolver el problema de agrupación, se puede elegir libremente la ordenación de las clases originales y, a continuación, buscar una partición contigua de esas clases que se ordenarán de modo que se minimice la distancia al cuadrado.

HANSEN, HURWITZ y MADOW, en su discusión de este problema, presentan el caso de una distribución de frecuencias de familias de Atlanta, que se agruparon en diez clases de niveles de renta. El problema que resolvieron consistió en combinar las diez

clases en tres estratos, de modo que la estimación de la renta media para todas las familias, basada sobre una muestra estratificada tuviera una pequeña varianza. La atención se fijó en las diversas combinaciones posibles de las clases originales en estratos, suponiendo una colocación proporcional de números de muestras entre los tres estratos y muestreo aleatorio dentro de cada estrato. Para hallar la agrupación óptima es suficiente calcular los valores de D para cada una de las 36 posibles particiones contiguas de diez elementos en tres grupos, y seleccionar aquella que ofrezca el menor (mínimo) valor de D.

Para algunos pequeños problemas sin restricciones, $K \leq 20$ y $G \leq 5$, se puede obtener la solución por enumeración completa de todas las posibles particiones contiguas, calculando los valores de D y seleccionando aquella que presente el menor valor, y en otros casos es posible obtener la solución, mediante una simple inspección ocular de la distribución de frecuencias de a_i, ordenando según sus magnitudes.

II. EL PROBLEMA CON RESTRICCIONES:

Del problema de agrupación anterior, sin restricciones, se pasa al problema CON RESTRICCIONES si al conjunto de particiones de K elementos en G subconjuntos admisibles para una solución, se añade alguna o algunas restricciones "a priori" y se exige, además, que los subconjuntos sean mutuamente exclusivos y exhaustivos.

Muchos de los problemas prácticos serán del tipo "con restricciones", ya que el investigador deseará añadir, o bien, ciertos conocimientos previos o factores de conveniencia a las condiciones de la agrupación.

Como caso interesante, aunque muy particular, de la multitud de problemas con restricciones que se pueden presentar, se puede considerar el que presenta una ordenación completa "a priori" de dos elementos, que es diferente del orden numérico de los a_i y que se debe a WALLIS y ROBERTS, quienes en la discusión de series de tiempo, presentan un ejemplo de cambio de niveles del lago Michigan - Huron, a lo largo de noventa y seis años.

En su gráfica sugieren ciertas épocas en las que el nivel era alto y otras, en las que dicho nivel era bajo, aunque sin regularidad o periodicidad. Supongamos que se deseara definir G épocas, tales que la variación del nivel del lago, dentro de las épocas, definido por la distancia cuadrada D, sea minimizado. Naturalmente, se exigirá que cada época comprenda únicamente años consecutivos: esta es una ordenación "a priori". Entonces, resulta un problema con restricciones en el que todos los pesos w_i son iguales a 1. La solución debe ser una partición contigua según la ordenación en el tiempo y no necesariamente según la ordenación de los niveles.

De este modo se han considerado tres problemas de interés:

a) Conglomeración, en general.

b) Conglomeración con la restricción de contigüidad sobre el terreno, como es, por ejemplo, el problema de la regionalización.

c) Problema de periodización que corresponde a la restricción de contigüidad con el tiempo.

LA DISTANCIA, δ, ENTRE CONGLOMERADOS

En el capítulo I hemos hablado del método de conglomeración δ que hemos sugerido. Dicho método utiliza la nueva distancia, que hemos llamado δ, entre conglomerados, y que esponemos a continuación:

Sean $A = \{a_1, a_2, \ldots, a_h\}$ y $B = \{b_1, b_2, \ldots, b_k\}$ dos conglomerado o clases constituídos, respectivamente, por los elementos $a_i, i = 1, \ldots h$ y $b_j, j = 1, \ldots k$, y sea d la métrica asociada al conjunto de OTUS.

DEFINICION: La distancia δ entre los conglomerados A y B es la siguiente:

$$\delta(A,B) = \frac{\max\{d(a_i,b_i)\} + \min\{d(a_i,b_i)\}}{2} \qquad \begin{matrix} i = 1,2,\ldots h \\ j = 1,2,\ldots k \end{matrix}$$

La distancia δ entre conglomerados se puede utilizar en el proceso de la conglomeración, una vez que entre los OTUS ha sido definida una métrica d.

Esta distancia puede considerarse como la aplicación a conglomerados o conjuntos de la noción de centro-recorrido (mid-range) que se aplica a distribuciones unidimensionales de frecuencia o probabilidad con recorrido finito.

CAPITULO IV

EVALUACION Y COMPARACION DE LAS TECNICAS DE CONGLOMERACION

PROCEDIMIENTOS Y COEFICIENTES PARA EVALUAR LAS TECNICAS DE CONGLOMERACION

Vamos a abordar, en este capítulo, los aspectos del análisis de conglomerados menos trabajados y, por tanto, con menor número de resultados.

Existen, en el análisis de conglomerados, multitud de problemas por resolver. Expondremos algunos de ellos y ofreceremos una revista de los resultados más destacados, hasta el presente, así como el procedimiento que hemos seguido para comparar métricas mediante el coeficiente de correlación cofenético a través de los valores cofenéticos de los correspondientes dendrogramas.

Cuando se quiere conglomerar un conjunto de elementos, el primer paso que se ha de dar es elegir una medida de similitud o disimilitud y, a continuación, seleccionar un método de análisis de conglomerados.

Y esa elección ha de hacerse sabiendo que:

1.- Existen tantas alternativas que es muy difícil decidir qué medida y qué método resolverá mejor un problema particular.

2.- Diferentes criterios de conglomeración aplicados al mismo conjunto de elementos pueden producir resultados distintos.

3.- Por otro lado, para un problema determinado, existe un gran número de diferentes métodos que darán, subs tancialmente los mismos resultados, mientras que, quizás, otros pocos proporcionarán resultados diferentes.

4.- Cada criterio de conglomeración está "predispuesto" a encontrar tipos particulares de conglomerados y puede darse el caso de que distorsione los resultados hacia ese ideal.

5.- Determinados métodos reaccionan a algunas características fundamentales de un problema miestras que hay otras que detectan características especiales, produciendose resultados distintos.

6.- En los datos pueden existir estructuras subyacientes que pasan inadvertidas a muchos métodos de conglomeración.

7.- Los conocimientos teóricos que se poseen son insuficientes para explicar satisfactoriamente, en todos los casos, la manera en la que los resultados de diversos métodos de conglomeración coinciden o difieren al ser aplicados a distintos conjuntos de datos.

Junto a las anteriores afirmaciones y a otras muchas - que pudieran hacerse, fruto de los trabajos en este campo, existen multitud de posibles preguntas, algunas de las cuales ofrecemos a continuación, y cuyas contestaciones intentan (las hay aún sin contestar, lo que hace atractivo este campo de investigación) paliar muchos de los problemas del análisis de conglomerados.

SNEATH y SOKAL (112) nos ofrecen algunas de tales preguntas.

a.- ¿Qué es una buena clasificación?

b.- ¿Es mejor la clasificación A que la B?

c.- ¿Son similares dos clasificaciones?¿en qué medida?

d.- ¿Es importante la estructura en un conjunto de OTUS?

e.- ¿Cual es la diferencia que existe entre los conglomerados obtenidos mediante procedimientos distintos?

A la vista de lo anteriormente dicho ¿cómo podemos abordar la cuestión de la evaluación de métodos alternativos de conglomeración?

Cada método nos ofrece una solución, y "a priori" todas las soluciones son igualmente buenas a menos que, claro está, las propiedades conocidas de un determinado método de conglomeración nos proporcionen alguna información adicional a la solución.

Por tanto, parece razonable, en una primera etapa, comenzar la tarea de la evaluación de los métodos de conglomeración tratando de establecer sus características fundamentales y a continuación, llevar a la práctica un procedimiento que se utiliza con frecuencia, y que consiste en resolver problemas de los que se conoce la solución.

Hechas las anteriores consideraciones, hemos de concluir que es necesario conseguir un procedimiento sistemáticos que rela-

cione las posibilidades de los métodos de conglomeración con las características fundamentales de los problemas a resolver.

Como afirma ANDERBERG (2) si fuera posible encontrar un conjunto de conceptos importantes que describan los problemas y otro conjunto que describa métodos, entonces podrían ser alcanzadas una variedad de importantes capacidades:

1.- Podrían ser construidas nuevas estrategias de análisis.

2.- Aspectos de coincidencia y de distinción entre métodos podrían ser identificados para ayudar a minimizar el número de particiones redundantes obtenidas al conglomerar un solo conjunto de datos.

3.- Problemas característicos que están fuera del alcance de los métodos de conglomeración podrían ser identificados y ayudar en la orientación de investigaciones que busquen nuevos métodos.

4.- El progreso podría hacerse en el esfuerzo para distinguir entre el poder de un método para revelar la estructura de un conjunto de datos o su tendencia a imponer una estructura.

Una posible aproximación para descubrir estas dimensiones conceptuales es volver el análisis de conglomerados sobre sí mismo y conglomerar los resultados obtenidos aplicando métodos provechosos a conjuntos de datos especialmente construidos.

Las similitudes y diferencias entre varios métodos de conglomeración pueden identificarse comparando los resultados obtenidos al conglomerar conjuntos de datos de características conocidas, y las características de diversos conjuntos de datos pueden descubrirse conglomerándolos con métodos de propiedades conocidas.

PROCEDIMIENTOS Y COEFICIENTES PARA EVALUAR LAS TECNICAS DE CONGLOMERACION

1.- El primer método que se obtuvo, en la historia de la taxonomía numérica, y que es debido a SOKAL y ROHLF en 1962, es el de las CORRELACIONES COFENETICAS. Este método asigna un VALOR COFENETICO a la similitud entre todo par de OTUS de un dendrograma y da lugar a una matriz de tales valores para todo conjunto de OTUS.

"El valor cofenético C_{jk} entre dos OTUS j y k es la similitud máxima S_{jk} (o disimilitud mínima d_{jk}) entre los dos OTUS del dendrograma". SNEATH y SOKAL (112).

Se calcula un coeficiente de correlación momento-producto entre los elementos S_{jk} de la matriz de similitud - original, S, y los valores cofenéticos C_{jk} de la matriz C. - Este coeficiente de correlación cofenético es una medida de la concordancia entre los valores de similitud del fenograma y los de la matriz original. Se ha observado que varía de 0,6 a 0,95 dependiendo del método que origina el fenograma y de - la estructura natural de los OTUS a clasificar.

Es un método muy popular entre biólogos y zoólogos, que lo utilizan como una medida de la bondad de ajuste entre la clasificación jerárquica y la matriz de similitud original.

Naturalmente, este importantísimo coeficiente (y el método) ha sido sometido a toda clase de pruebas y se ha trajado mucho con él y sobre él.

FARRIS (2) investiga métodos de conglomeración que tratan de maximizar el mencionado coeficiente y concluya que tales métodos son excesivamente sensitivos para el tamaño del conglomerado. Sin embargo, el coeficiente de correlación cofenético que es un útil utensilio evaluativo, falla como criterio algorítmico (2), o sea, que dicho coeficiente no sirve para clasificar o conglomerar, sino para comparar las clasificaciones o conglomeraciones.

BORKO (2) utilizó virtualmente este mismo esquema de evaluación para comparar diferentes clasificaciones derivadas de un único conjunto de datos unitarios

2.- HARTIGAN prefirió caracterizar la comparación de matrices de similitud utilizando distancias mejor que correlaciones:

En particular, si $S_k(i,j)$ es el valor de similitud(o medida de asociación) entre las entidades i y j expresado en la k-ésima matriz de similitud, entonces la distancia entre las matrices de similitud p y q es:

$$R(S_p,S_q) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} W(i,j) \left[S_p(i,j) - S_q(i,j)\right]^2$$

siendo n el número de entidades o elementos y W (i,j) una función de ponderación.

Los índices de los sumatorios i y j corresponden a filas y columnas, respectivamente, de la más pequeña matriz de similitud.

3.- Una medida de la DISTORSION (el complemento de la concordan-

cia representada por el coeficiente de correlación cofenético) nos la proporciona una familia de medidas; que es debida a - JARDINE y SIBSON (ver 70). Este coeficiente está dado por

$$\hat{\Delta}_{\mu} = \frac{\left[\sum_{jk} (U_{jk}-C_{jk})^{1/\mu}\right]^{\mu}}{\left[\sum_{jk} U_{jk}^{1/\mu}\right]^{\mu}}$$

donde U_{jk} es el coeficiente de disimilitud entre los OTUS j y k ($j \neq k$), C_{jk} es su valor cofenético y μ es un coeficiente arbitrario tal que $0 \leq \mu \leq 1$. Variando el coeficiente es posible acentuar las mayores o menores diferencias entre las disimilitudes y los valores cofenéticos respectivamente.

4.- MEZZICH (116) en un interesantísimo estudio abordó y resolvió el problema de la evaluación de procedimientos de conglomeración. Construyó, para ello, tres índices: una medida de validez externa, una medida de validez interna y una medida de - replicabilidad.

El mencionado autor midió el criterio de validez - externa obteniendo el porcentaje de concordancia de las predicciones de los expertos y los resultados de los procedimientos de conglomeración; esto puede ser llevado a cabo con una tabla de contingencia.

La medida del criterio de validez interna es el - coeficiente de correlación cofenético.

La medida de la REPLICABILIDAD o ESTABILIDAD es - también, esencialmente, un coeficiente de correlación. Los

datos iniciales se dividen aleatoriamente en dos conjuntos de igual número de elementos. Para cada uno de los dos conjuntos de datos se obtiene la configuración de conglomeración correspondiente mediante el procedimiento en cuestión. De cada una de las dos configuraciones de conglomeración se deriva una matriz de similitud y es entonces cuando se calcula un coeficiente de correlación a partir de cada una de ellas.

5.- A continuación exponemos el proceso referido, SOLOMON, para conglomerar dos técnicas de conglomeración sobre la base de una matriz de similitud.

Supongamos que consideramos una pareja cualquiera de técnicas de conglomeración, A y B que ya han conglomerado un determinado conjunto, el constituido por la mitad de los datos iniciales; hemos obtenido por tanto 2 configuraciones de similitud que darán lugar a sendas matrices de similitud.

Se puede calcular un coeficiente de correlación de una manera similar al coeficiente de correlación cofenético obteniendo el coeficiente de correlación momento producto sobre todos los pares de elementos correspondientes en las dos últimas matrices de similitud. Como este proceso puede repetirse para la mitad de las otras tres bases de datos obtendremos cuatro coeficientes de correlación diferentes para el mismo elemento si construimos una matriz de correlación producida por los procedimientos de conglomeración.

Utilicemos la medida de las cuatro correlaciones y

tendremos una matriz de correlación en la que las filas y columnas son los diferentes procedimientos de conglomeración. Una matriz similar se construye a partir de la otra mitad de cada una de las cuatro bases de datos.

Como el procedimiento de la UNION COMPLETA es el mejor algoritmo de conglomeración para las cuatro bases de datos empleadas anteriormente, lo aplicamos a cada una de las dos matrices de correlación de conglomerado y observamos los resultados; a continuación descubriremos que la UNION COMPLETA y el CRITERIO DE OPTIMACION de la COVARIANZA son las técnicas que se encuentran en un conglomerado y juntas con los procedimientos de conglomeración denominados ISODATA y de KING forman un conglomerado de cuatro procedimientos de conglomeración.

6.- Una comparación de los métodos de SOKAL y MIGHENER, EDWARDS y CAVALLI-SFORZA y WILLIAMS y LAMBERT nos la ofrece, GOWER ( ) en un interesantísimo trabajo.

RAND presenta criterios objetivos para comparar dos conglomeraciones diferentes sobre el mismo conjunto de datos.

FISHER y VAN NESS definen condiciones de admisibilidad que serían deseables bajo ciertas circunstancias y comparan con ellas varios métodos de conglomeración. Dichas condiciones de admisibilidad son formuladas de manera que rechazan aquellos métodos que originen "malas conglomeraciones". Sin embargo, podría darse el caso de que rechazaran algún procedimiento con resultados razonables.

HARTIGAN, JOHNSON y JARDINE y SIBSON discuten y estudian estas condiciones.

7.- Para la comparación de distancias euclideas, GOWER (23) - sigue el siguiente procedimiento:

Cualquier conjunto de distancias euclideas d_{ij} puede ser representado por puntos X_i tales que las distancias - $(X_i,X_j) = d_{ij}$. Para comparar d_{ij} con d^*_{ij} se gira el conjunto de puntos X_i relativo a X^*_i hasta que $M^2 = \sum_{i=1}^{n} \Delta^2(X_i,X^*_i)$ sea mínimo. El coeficiente de distorsión M^2, que puede ser -- utilizado como una distancia, es aplicable incluso en el caso de que las d^*_{ij} sean distancias ultramétricas y, por tanto, M^2 podría utilizarse como un criterio de conglomeración jerárquica.

8.- GORDON (54) expone un algoritmo para el estudio de dendrogramas equivalentes de orden local maximal y que está basado en la siguiente definición de SIBSON (109):

Si T_1 y T_2 son dos dendrogramas, se dice que son - equivalentes de orden local si

$$\hat{d}^{(1)}_{ij} \leq \hat{d}^{(1)}_{ik} \quad \text{y si y solo si}$$

$\hat{d}^{(2)}_{ij} \leq \hat{d}^{(2)}_{ik}$ para todos los elementos i,j,k, siendo $\hat{d}^{(k)}_{ij}$ el más bajo nivel al que los objetos i y j pertenecen al mismo grupo en el dendrograma k; en el mismo trabajo también se habla de la equivalencia de orden global.

También se han efectuado estudios comparativos utilizando los métodos geométricos, en los que cada objeto o elemento es representado por un punto de un espacio de baja dimensión, entre ellos el método del, denominado por su rigidez, análisis de Procust o Procustes analysis.

Como afirma GORDON (54), "probablemente, los más fructíferos estudios comparativos son los que combinan conglomeraciones con métodos geométricos.

9.- JARDINE y SIBSON (70) en su estudio sobre la evaluación de conglomeraciones, comparan, en primer lugar, coeficientes de disimilitud, entradas (INPUTS) y salidas (OUTPUTS), con objeto de conocer cómo ha funcionado un método de conglomeración en un determinado caso y después comparan dos coeficientes de disimilitud.

Si d y D(d) son, respectivamente, el coeficiente de similitud de entradas (inputs) y D(d) de salidad (outputs), los representan mediante puntos de un espacio euclideo de dimensión $\frac{1}{2}\left[n(n-1)\right]$, siendo n el conjunto de OTUS a clasificar.

Aunque se puede trabajar con la métrica euclidea, plantean el uso de las tres métricas siguientes:

Si d_1 y d_2 son dos coeficientes de disimilitud

1.- $\delta_0(d_1,d_2) = \max\left\{ \left| d_1(x,y) - d_2(x,y) \right| \;/\; x,y \in X \right\}$

2.- $\delta_{\frac{1}{2}}(d_1,d_2) = \left\{ \sum \left[d_1(x,y) - d_2(x,y) \right]^2 \right\}^{\frac{1}{2}}$

3.- $\delta_1(d_1,d_2) = \sum \left| d_1(x,y) - d_2(x,y) \right|$

teniendo en cuenta que en los dos últimos casos, los sumatorios se toman sobre los $\frac{1}{2}$ n (n-1) subconjuntos de los elementos de X.

Para evitar el factor de dimensionalidad es conveniente dividir por un factor normalizante y considerar:

$$\bar{\delta}_0 = \delta_0$$

$$\bar{\delta}_{\frac{1}{2}} = \left[\tfrac{1}{2}\, n\, (n-1)\right]^{-\frac{1}{2}}\ \delta_{\frac{1}{2}}$$

$$\bar{\delta}_1 = \left[\tfrac{1}{2}\, n\, (n-1)\right]^{-1}\ \delta_1$$

Tanto las primeras métricas como las segundas, son métricas sobre el conjunto C(X), de todos los coeficientes de disimilitud sobre X.

Ahora bien, para comparar métricas es necesario tam bién que dicha comparación sea independiente de las escalas d medida en ambos coeficientes de disimilitud, para lo cual hay que normalizar d_1 y d_2 y, a continuación, utilizar la métrica conveniente para compararlas.

De esta forma se obtiene:

$$\delta_0^* (d_1, d_2) = \delta_0 \left[d_1 / S_0(d_1)\ ,\ d_2 / S_0(d_2)\right]$$

$$\delta_{\frac{1}{2}}^* (d_1, d_2) = \delta_{\frac{1}{2}} \left[d_1 / S_{\frac{1}{2}}(d_1)\ ,\ d_2 / S_{\frac{1}{2}}(d_2)\right]$$

$$\delta_1^* (d_1,d_2) = \delta_1 \left[d_1/S_1(d_1),\ d_2/S_1(d_2) \right],$$

siendo $S_0(d) = \delta_0(d,\bar{0})$

$S_{1/2}(d) = \delta_{1/2}(d,\bar{0})$

$S_1(d) = \delta_1(d,\bar{0})$ donde $\bar{0}$ es el coeficiente de disimilitud representado por el origen del espacio euclideo asociado, es decir, $d(x,y)=0 \ \forall \ x,y \in X$

Otro método de comparación de coeficientes de disimilitud, considera la diferencia entre d_1 y d_2 como el ángulo entre las rectas Od_1 y Od_2,

$$\alpha (d_1,d_2) = \cos^{-1} \left\{ \frac{\left[S_{1/2}(d_1)\right]^2 + \left[S_{1/2}(d_2)\right]^2 - \left[\delta_{1/2}(d_1,d_2)\right]^2}{2\, S_{1/2}(d_1)\, S_{1/2}(d_2)} \right\}$$

0.- El procedimiento que hemos seguido para comparar métricas, y que desarrollamos en el capítulo VI, consiste en lo siguiente:

Una vez definido un conjunto de OTUS y obtenidas las matrices de distancias, a comparar, entre dichos OTUS y definida una métrica δ^* entre las distancias anteriores conseguimos una matriz de distancias δ^* , a la que aplicamos diversos procedimientos de conglomeración, cada uno de los cuales da lugar a un dendrograma que le corresponde una matriz de valores cofenéticos. Las matrices anteriores las comparamos mediante el coeficiente de correlación cofenético de SOKAL, coeficiente que nos indicará la concordancia entre los dendrogramas mencionados.

CAPITULO V

APLICACION DE ALGUNOS METODOS DE CONGLOMERACION PARA LA CLASIFICACION DE NEUROPTERIS Y COMPARACION DE LOS RESULTADOS.

Este capítulo consta de dos partes estrechamente relacionadas pero muy bien diferenciadas.

La primera de ellas está dedicada a la aplicación, y ello se hace por primera vez, de los métodos de taxonomía numérica para la clasificación de ejemplares del género NEUROPTERIS, - PTERIDOSPERMEAS fósiles del Carbonífero Superior, resolviendo un problema planteado en Paleobotánica; entre los métodos que hemos aplicado figura el método , propuesto por nosotros, y que figura en el capítulo II.

La segunda parte consiste en la comparación, mediante el coeficiente de correlación cofenético, que hemos expuesto en el capítulo V, de los métodos utilizados en la primera parte.

CLASIFICACION DE PTERIDOSPERMAS

Como ya se ha indicado vamos a utilizar diversos métodos, enlace sencillo o distancia mínima, enlace completo o distancia máxima, método de la media y método δ para clasificar 13 ejemplares del género NEUROPTERIS.

La primera parte,todo método de clasificación, como es la elección del conjunto de OTUS a clasificar y la asignación de caracteres a cada OTU,ha sido realizada en el laboratorio de Paleobotánica de la Facultad de Ciencias Geológicas de la Universidad Complutense en un trabajo dirigido por el Dr. Talens, que fotografiaron 13 ejemplares de NEUROPTERIS (en Paris), de los que obtuvieron diversas medidas con objeto de clasificarlos; tal fue el material que nos proporcionaron.

Hay que hacer la consideración de que la adscripción específica en los fósiles, a falta de otro tipo de caracteres (posibilidad de reproducción, grupos serológicos, número de cromosomas, etc.) se hace estrictamente sobre caracteres morfológicos, hablándose, incluso, de "parataxones de forma" para indicar los géneros y especies definidos en fósiles.

Se nos proporcionaron los siguientes datos correspondientes a los 13 ejemplares mencionados:

A — (3, 2.1, 1.2, 2.3)

B — (3.2, 2.2, 1.4, 2.3)

C — (2.3, 1.3, 1.4, 1.8)

D — (2.6, 1.2, 1.2, 2)

E — (3.7, 1.3, 1.7, 2.8)

F — (3.9, 1.7, 1.2, 2.7)

G — (4.9, 1.4, 0.7, 4.1)

H — (6.6, 3.3, 2, 5.8)

I — (6.9, 3.7, 2.1, 6.1)

J — (5.7, 2.4, 1.1, 5)

K — (4.8, 1.6, 1.4, 3.9)

L — (4.6, 2.2, 1.8, 3.8)

M — (7.8, 2.6, 1.4, 6.6)

es decir a cada elemento se le ha asociado 4 caracteres morfológicos.

El primer paso que hemos dado ha sido construir la matriz de distancias asociada a la partión inicial:

$P_0: \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}, \{G\}, \{H\}, \{I\}, \{J\}, \{K\}, \{L\}, \{M\}\}$, para lo que hemos utilizado la distancia euclidea:

$$d(X_i, X_j) = \sqrt{\sum_{i=1}^{4} (x_i - y_i)^2}$$

y hemos obtenido la siguiente matriz:

000200

	A	B	C	D	E	F	G	H	I	J	K	L	M
	3	3,2	2,3	2,6	3,7	3,9	4,9	6,6	6,9	5,7	4,8	4,6	7,8
	2,1	2,2	1,3	1,2	1,3	1,7	1,4	3,3	3,7	2,4	1,6	2,2	2,6
	1,2	1,4	1,4	1,2	1,7	1,2	0,7	2	2,1	1,1	1,4	1,8	1,4
	2,3	2,3	1,8	2	2,8	2,7	4,1	5,8	6,1	5	3,9	3,8	6,6
A 3 - 2,1 - 1,2 - 2,3	0	0,3	1,19	1,04	1,27	1,06	2,69	5,22	5,74	3,82	2,46	2,07	6,46
B 3,2 - 2,2 - 1,4 - 2,3		0	1,36	1,22	1,1	0,96	2,69	5,03	6,2	3,69	2,34	2,09	6,3
C 2,3 - 1,3 - 1,4 - 1,8			0	0,42	1,74	1,88	3,54	5,98	6,77	4,76	3,27	3,2	7,41
D 2,6 - 1,2 - 1,2 - 2				0	1,45	1,55	3,16	5,95	6,91	4,47	2,94	2,93	7,84
E 3,7 - 1,3 - 1,7 - 2,8					0	0,67	2,03	4,63	5,24	3,22	1,61	1,62	5,61
F 3,9 - 1,7 - 1,2 - 2,7						0	1,81	4,48	5,03	3	1,51	1,51	5,59
G 4,9 - 1,4 - 0,7 - 4,3							0	2,06	3,9	1,61	1,78	1,42	4,07
H 6,6 - 3,3 - 2 - 5,8								0	0,53	1,75	3,17	3,04	1,71
I 6,9 - 3,7 - 2,1 - 6,1									0	2,31	3,7	3,59	1,66
J 5,7 - 2,4 - 1,1 - 5										0	1,65	1,78	2,66
K 4,8 - 1,6 - 1,4 - 3,9											0	0,75	4,15
L 4,6 - 2,2 - 1,8 - 3,8												0	4,28
M 7,8 - 2,6 - 1,4 - 6,6													0

que una vez normalizada se convierte en la matriz:

000202

	A	B	C	D	E	F	G	H	I	J	K	L	M
	3	3,2	2,3	2,6	3,7	3,9	4,9	6,6	6,9	5,7	4,8	4,6	7,8
	2,1	2,2	1,3	1,2	1,3	1,7	1,4	3,3	3,7	2,4	1,6	2,2	2,6
	1,2	1,4	1,4	1,2	1,7	1,2	0,7	2	2,1	1,1	1,4	1,8	1,4
	2,3	2,3	1,8	2	2,8	2,7	4,1	5,8	6,1	5	3,9	3,8	6,6
A 3 - 2,1 - 1,2 - 2,3	0	0,03	0,15	0,13	0,16	0,13	0,34	0,66	0,73	0,48	0,31	0,26	0,82
B 3,2 - 2,2 - 1,4 - 2,3		0	0,17	0,15	0,14	0,12	0,34	0,64	0,79	0,47	0,29	0,26	0,80
C 2,3 - 1,3 - 1,4 - 1,8			0	0,05	0,22	0,23	0,45	0,76	0,86	0,60	0,41	0,40	0,94
D 2,6 - 1,2 - 1,2 - 2				0	0,18	0,19	0,40	0,75	0,88	0,57	0,37	0,37	1
E 3,7 - 1,3 - 1,7 - 2,8					0	0,08	0,25	0,59	0,66	0,41	0,20	0,20	0,71
F 3,9 - 1,7 - 1,2 - 2,7						0	0,23	0,57	0,64	0,38	0,19	0,19	0,71
G 4,9 - 1,4 - 0,7 - 4,3							0	0,391	0,49	0,20	0,22	0,18	0,51
H 6,6 - 3,3 - 2 - 5,8								0	0,06	0,22	0,40	0,38	0,21
I 6,9 - 3,7 - 2,1 - 6,1									0	0,29	0,47	0,45	0,21
J 5,7 - 2,4 - 1,1 - 5										0	0,21	0,22	0,33
K 4,8 - 1,6 - 1,4 - 3,9											0	0,09	0,52
L 4,6 - 2,2 - 1,8 - 3,8												0	0,54
M 7,8 - 2,6 - 1,4 - 6,6													0

A partir de la matriz anterior hemos comenzado a aplicar los cuatro métodos mencionados; el proceso seguido, que está explicado en el capítulo II, y los resultados que en las diferentes etapas hemos obtenido figuran ordenadamente a continuación:

I. METODO DE LA DISTANCIA MINIMA:

Como el menor elemento de la matriz es 0,03, los elementos A y B se fusionan para formar el primer conglomerado al nivel 0,03.

Hemos de obtener, a continuación, las nuevas distancias entre dicho conglomerado $\{A, B\}$ y los restantes elementos:

$$d\left[(A,B),C\right] = \text{mín}\ \ d(A,C),\ d(B,C)\ \ = \text{mín}\ \ 0,15,\ 0,15 = 0,15$$

$$d\left[(A,B),D\right] = 0,15$$

$$d\left[(A,B),E\right] = 0,14$$

$$d\left[(A,B),F\right] = 0,12$$

$$d\left[(A,B),G\right] = 0,34$$

$$d\left[(A,B,),H\right] = 0,64$$

$$d\left[(A,B),I\right] = 0,73$$

$$d\left[(A,B),J\right] = 0,47$$

$$d\left[(A,B),K\right] = 0,29$$

$$d\left[(A,B),L\right] = 0,26$$

$$d\left[(A,B),M\right] = 0,80$$

luego la nueva matriz de distancias es:

LUEGO LA NUEVA MATRIZ ES:

	A-B	C	D	E	F	G	H	I	J	K	L	M
A-B	0	0,15	0,15	0,14	0,12	0,34	0,64	0,73	0,47	0,29	0,26	0,80
C		0		0,22	0,23	0,45	0,76	0,86	0,60	0,41	0,40	0,94
D			0	0,18	0,19	0,40	0,75	0,88	0,57	0,37	0,37	1
E				0	0,	0,25	0,59	0,66	0,41	0,20	0,20	0,71
F					0	0,23	0,57	0,64	0,38	0,19	0,19	0,71
G						0	0,391	0,49	0,20	0,22	0,18	0,51
H							0	0,06	0,22	0,40	0,38	0,21
I								0	0,29	0,47	0,45	0,21
J									0	0,21	0,22	0,33
K										0	0,09	0,52
L											0	0,54
M												0

NUEVO GRUPO: C-D AL NIVEL 0,05

LUEGO LA NUEVA MATRIZ ES:

	A-B	C-D	E	F	G	H	I	J	K	L	M
A-B	0	0,13	0,14	0,12	0,34	0,64	0,73	0,47	0,29	0,26	0,80
C-D		0	0,18	0,19	0,40	0,75	0,86	0,57	0,37	0,37	0,94
E			0	0,08	0,25	0,59	0,66	0,41	0,20	0,20	0,71
F				0	0,23	0,57	0,64	0,38	0,19	0,19	0,71
G					0	0,391	0,49	0,20	0,22	0,18	0,51
H						0	0,06	0,22	0,40	0,38	0,21
I							0	0,29	0,47	0,45	0,21
J								0	0,21	0,22	0,33
K									0	0,09	0,52
L										0	0,54
M											0

SE FORMA UN NUEVO GRUPO E-F AL NIVEL 0,06

Y LA CORRESPONDIENTE MATRIZ ES:

	A-B	C-D	H-I	E	F	G	J	K	L	M
A-B	0	0,13	0,64	0,14	0,12	0,34	0,47	0,29	0,26	0,80
C-D		0	0,75	0,18	0,19	0,40	0,41	0,20	0,20	0,71
H-I			0	0,59	0,57	0,391	0,22	0,40	0,38	0,21
E				0	0,08	0,25	0,41	0,20	0,20	0,71
F					0	0,23	0,38	0,19	0,19	0,71
G						0	0,20	0,22	0,18	0,51
J							0	0,21	0,22	0,33
K								0	0,09	0,52
L									0	0,54
M										0

NUEVO GRUPO: E-F AL NIVEL 0,08

Y LA NUEVA MATRIZ ES:

	A-B	C-D	H-I	E-F	G	J	K	L	M
A-B	0	0,13	0,64	0,12	0,34	0,47	0,29	0,26	0,80
C-D		0	0,75	0,18	0,40	0,41	0,20	0,20	0,71
H-I			0	0,57	0,39	0,22	0,40	0,38	0,21
E-F				0	0,23	0,38	0,19	0,19	0,71
G					0	0,20	0,22	0,18	0,51
J						0	0,21	0,22	0,33
K							0	0,09	0,52
L								0	0,54
M									0

COMO EL ELEMENTO MAS PEQUENO DE LA MATRIZ ES 0,09

A ESE NIVEL SE FORMA UN NUEVO GRUPO : K-L

Y LA CORRESPONDIENTE MATRIZ ES :

	A-B	C-D	H-I	E-F	K-L	G	J	M
A-B	0	0,13	0,64	0,12	0,26	0,34	0,47	0,8
C-D		0	0,75	0,18	0,20	0,40	0,41	0,71
H-I			0	0,57	0,38	0,39	0,22	0,21
E-F				0	0,19	0,23	0,38	0,71
K-L					0	0,18	0,21	0,52
G						0	0,20	0,51
J							0	0,33
M								0

$[(A\text{-}B),(E\text{-}F)]$ = 0,12 QUE ES LA MAS PEQUEÑA, LUEGO YA TENEMOS UN NUEVO GRUPO: $[(A\text{-}B),(E\text{-}F)]$ AL NIVEL 0,12

Y LA CORRESPONDIENTE MATRIZ ES:

	$[(A\text{-}B),(E\text{-}F)]$	C-D	H-I	K-L	G	J	M
$[(A\text{-}B),(E\text{-}F)]$	0	0,13	0,57	0,19	0,23	0,38	0,71
C-D		0	0,75	0,20	0,40	0,41	0,71
H-I			0	0,38	0,39	0,22	0,21
K-L				0	0,18	0,21	0,52
G					0	0,20	0,51
J						0	0,33
M							0

UN NUEVO GRUPO $\{[(A\text{-}B),(E\text{-}F)],(C\text{-}D)\}$ AL NIVEL 0,13

Y LA NUEVA MATRIZ ES:

	$[(A\text{-}B),(E\text{-}F)],(C\text{-}D)$	H-I	K-L	G	J	M
$[(A\text{-}B),(E\text{-}F)],(C\text{-}D)$	0	0,57	0,19	0,23	0,38	0,71
H-I		0	0,38	0,39	0,22	0,21
K-L			0	0,18	0,21	0,52
G				0	0,20	0,51
J					0	0,33
M						0

EL NUEVO GRUPO, AHORA ES: [(K-L), G] AL NIVEL 0,18

Y LA NUEVA MATRIZ ES:

	{[(A-B),(E-F)],(C-D)}	[(K-L),G]	H-I	J	M
{[(AB),(E-F)],(C-D)}	0	0,19	0,57	0,38	0,71
[(K-L),G]		0	0,38	0,20	0,51
H-I			0	0,22	0,21
J				0	0,33
M					0

NUEVO GRUPO {[(A-B),(E-F)],(C-D)},[(K-L),G]

AL NIVEL 0,19 Y LA MATRIZ ES:

	{[(A-B),(E-F)],(C-D)},[(K-L),G]	H-I	J	M
{[(AB),(E-F)],(C-D)},[(K-L),G]	0	0,38	0,20	0,51
H-I		0	0,22	0,21
J			0	0,33
M				0

NUEVO GRUPO: {[(A-B),(E-F)],(C-D)},[(K-L),G], J
(G*)

LA NUEVA MATRIZ ES:

	(G*-J)	H-I	M
(G*-J)	0	0,22	0,33
H-I		0	0,21
M			0

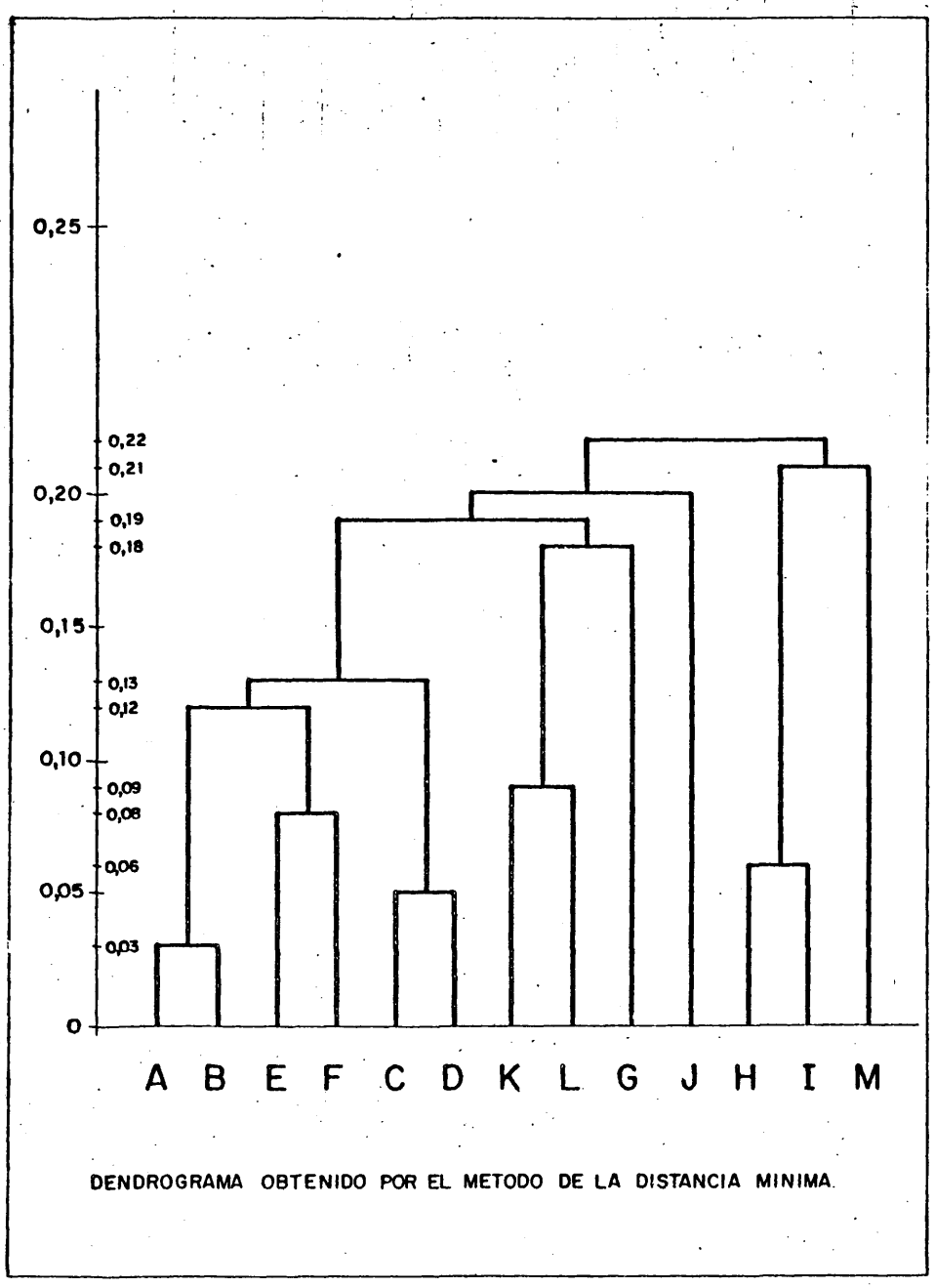
LUEGO EL NUEVO GRUPO ES [(H-I), M]

AL NIVEL 0,21 Y LA NUEVA MATRIZ ES:

	(G*-J)	[(H-I), M]
(G*-J)	0	0,22
[(H-I), M]		0

POR TANTO EL DENDROGRAMA CORRESPONDIENTE ES:

000213



DENDROGRAMA OBTENIDO POR EL METODO DE LA DISTANCIA MINIMA.

II. METODO DE ENLACE COMPLETO O DISTANCIA MAXIMA:

Como el elemento más pequeño de la matriz es 0,03, los elementos A y B se unen para formar el primer grupo (A,B).

Hallamos las distancias máximas de este grupo a los restantes elementos:

$$d\left[(A,B),C\right] = \text{máx}\left\{d_1(A,C),\ d_1(B,C)\right\} = 0,17$$

$$d\left[(A,B),D\right] = \text{máx}\left\{d_1(A,D),\ d_1(B,D)\right\} = 0,15$$

$$d\left[(A,B),E\right] = \text{máx}\left\{d_1(A,E),\ d_1(B,E)\right\} = 0,16$$

$$d\left[(A,B),F\right] = \text{máx}\left\{d_1(A,F),\ d_1(B,F)\right\} = 0,13$$

$$d\left[(A,B),G\right] = \text{máx}\left\{d_1(A,G),\ d_1(B,G)\right\} = 0,34$$

$$d\left[(A,B),H\right] = \text{máx}\left\{d_1(A,H),\ d_1(B,H)\right\} = 0,66$$

$$d\left[(A,B),I\right] = \text{máx}\left\{d_1(A,I),\ d_1(B,I)\right\} = 0,79$$

$$d\left[(A,B),J\right] = \text{máx}\left\{d_1(A,J),\ d_1(B,J)\right\} = 0,84$$

$$d\left[(A,B),K\right] = \text{máx}\left\{d_1(A,K),\ d_1(B,K)\right\} = 0,31$$

$$d\left[(A,B),L\right] = \text{máx}\left\{d_1(A,L),\ d_1(B,L)\right\} = 0,26$$

$$d\left[(A,B),M\right] = \text{máx}\left\{d_1(A,M),\ d_1(B,M)\right\} = 0,82$$

METODO: ENLACE COMPLETO O DISTANCIA MAXIMA (NEUROPTERIS)

COMO EL ELEMENTO MAS PEQUEÑO DE LA MATRIZ ES 0,03 LOS ELEMENTOS A Y B SE UNEN PARA FORMAR EL PRIMER GRUPO (A-B) HALLAMOS LAS DISTANCIAS MAXIMAS DE ESTE GRUPO A LOS RESTANTES ELEMENTOS:

$d[(A\text{-}B),C] = \text{MAXIMA}\{d_1(A\text{-}C), d_1(B\text{-}C)\} = 0{,}17$

$d[(A\text{-}B),D] = \text{MAXIMA}\{d_1(A\text{-}D), d_1(B\text{-}D)\} = 0{,}15$

$d[(A\text{-}B),E] = \text{MAXIMA}\{d_1(A\text{-}E), d_1(B\text{-}E)\} = 0{,}16$

$d[(A\text{-}B),F] = \text{MAXIMA}\{d_1(A\text{-}F), d_1(B\text{-}F)\} = 0{,}13$

$d[(A\text{-}B),G] = \text{MAXIMA}\{d_1(A\text{-}G), d_1(B\text{-}G)\} = 0{,}34$

$d[(A\text{-}B),H] = \text{MAXIMA}\{d_1(A\text{-}H), d_1(B\text{-}H)\} = 0{,}66$

$d[(A\text{-}B),I] = \text{MAXIMA}\{d_1(A\text{-}I), d_1(B\text{-}I)\} = 0{,}79$

$d[(A\text{-}B),J] = \text{MAXIMA}\{d_1(A\text{-}J), d_1(B\text{-}J)\} = 0{,}48$

$d[(A\text{-}B),K] = \text{MAXIMA}\{d_1(A\text{-}K), d_1(B\text{-}K)\} = 0{,}31$

$d[(A\text{-}B),L] = \text{MAXIMA}\{d_1(A\text{-}L), d_1(B\text{-}L)\} = 0{,}26$

$d[(A\text{-}B),M] = \text{MAXIMA}\{d_1(A\text{-}M), d_1(B\text{-}M)\} = 0{,}82$

LA NUEVA MATRIZ DE DISTANCIAS MAXIMAS ES:

	A-B	C	D	E	F	G	H	I	J	K	L	M
A-B	0	0,17	0,15	0,16	0,13	0,34	0,66	0,79	0,48	0,31	0,26	0,82
C		0	0,05	0,22	0,23	0,45	0,76	0,86	0,60	0,41	0,40	0,94
D			0	0,18	0,19	0,40	0,75	0,88	0,57	0,37	0,37	1
E				0	0,08	0,25	0,59	0,66	0,41	0,20	0,20	0,71
F					0	0,23	0,57	0,64	0,38	0,19	0,19	0,71
G						0	0,391	0,49	0,20	0,22	0,18	0,51
H							0	0,06	0,22	0,40	0,38	0,21
I								0	0,29	0,47	0,45	0,21
J									0	0,21	0,22	0,33
K										0	0,09	0,52
L											0	0,54
M												0

EL ELEMENTO MAS PEQUEÑO DE LA MATRIZ ANTERIOR ES 0,05 LUEGO LOS ELEMENTOS **C** Y **D** SE UNEN PARA FORMAR UN NUEVO GRUPO A DICHO NIVEL:

LA MATRIZ DE DISTANCIAS ES:

	A-B	C-D	E	F	G	H	I	J	K	L	M
A-B	0	0,17	0,16	0,13	0,34	0,66	0,79	0,48	0,31	0,26	0,82
C-D		0	0,22	0,23	0,45	0,76	0,88	0,60	0,41	0,40	1
E			0	0,08	0,25	0,59	0,66	0,41	0,20	0,20	0,71
F				0	0,23	0,57	0,64	0,38	0,19	0,19	0,71
G					0	0,391	0,49	0,20	0,22	0,18	0,51
H						0	0,06	0,22	0,40	0,38	0,21
I							0	0,29	0,47	0,45	0,21
J								0	0,21	0,22	0,33
K									0	0,09	0,52
L										0	0,54
M											0

EL MENOR ELEMENTO DE LA MATRIZ ES 0,06 LUEGO LOS ELEMENTOS H E I FORMAN UN NUEVO GRUPO, LA NUEVA MATRIZ DE DISTANCIAS ES:

	A-B	C-D	H-I	E	F	G	J	K	L	M
A-B	0	0,17	0,79	0,16	0,13	0,34	0,48	0,31	0,26	0,82
C-D		0	0,88	0,22	0,23	0,45	0,60	0,41	0,40	1
H-I			0	0,66	0,64	0,49	0,29	0,47	0,45	0,21
E				0	0,08	0,25	0,41	0,20	0,20	0,71
F					0	0,23	0,38	0,19	0,19	0,71
G						0	0,20	0,22	0,18	0,51
J							0	0,21	0,22	0,33
K								0	0,09	0,52
L									0	0,54
M										0

ELEMENTO MAS PEQUEÑO DE LA MATRIZ ANTERIOR 0,08

NUEVO GRUPO E-F

MATRIZ DE DISTANCIAS MAXIMAS:

	A-B	C-D	H-I	E-F	G	J	K	L	M
A-B	0	0,17	0,79	0,16	0,34	0,48	0,31	0,26	0,82
C-D		0	0,88	0,23	0,45	0,60	0,41	0,40	1
H-I			0	0,66	0,49	0,29	0,47	0,45	0,21
E-F				0	0,25	0,41	0,20	0,20	0,71
G					0	0,20	0,22	0,18	0,51
J						0	0,21	0,22	0,33
K							0	0,09	0,52
L								0	0,54
M									0

ELEMENTO MAS PEQUEÑO DE LA MATRIZ ANTERIOR 0,09

NUEVO GRUPO K-L

MATRIZ DE DISTANCIAS MAXIMAS:

	A-B	C-D	H-I	E-F	K-L	G	J	M
A-B	0	0,17	0,79	0,16	0,31	0,34	0,48	0,82
C-D		0	0,88	0,23	0,41	0,45	0,60	1
H-I			0	0,66	0,47	0,49	0,29	0,21
E-F				0	0,20	0,25	0,41	0,71
K-L					0	0,22	0,22	0,54
G						0	0,20	0,51
J							0	0,33
M								0

MENOR ELEMENTO DE LA MATRIZ ANTERIOR 0,16

NUEVO GRUPO [(A-B),(E-F)]

	[(A-B),(E-F)]	C-D	H-I	K-L	G	J	M
[(A-B),(E-F)]	0	0,23	0,79	0,31	0,34	,48	0,82
C-D		0	0,88	0,41	0,45	0,60	1
H-I			0	0,47	0,49	0,29	0,21
K-L				0	0,22	0,22	0,54
G					0	0,20	0,51
J						0	0,33
M							0

ELEMENTO MAS PEQUEÑO DE LA MATRIZ ANTERIOR 0,20

NUEVO GRUPO G-J

MATRIZ DE DISTANCIAS MAXIMAS:

	[(A-B),(E-F)]	C-D	H-I	K-L	G-J	M
[(A-B),(E-F)]	0	0,23	0,79	0,31	0,48	,82
C-D		0	0,88	0,41	0,60	1
H-I			0	0,47	0,49	,21
K-L				0	0,22	,54
G-J					0	0,51
M						0

ELEMENTO MAS PEQUENO DE LA MATRIZ ANTERIOR 0,21

NUEVO GRUPO [(H-I) M]

MATRIZ DE DISTANCIAS:

	[(A-B),(E-F)]	[(H-I),M]	C-D	K-L	G-J
[(A-B),(E-F)]	0	0,82	0,23	0,31	0,48
[(H-I), M]		0	1	0,54	0,51
C-D			0	0,41	0,60
K-L				0	0,22
G-J					0

ELEMENTO MENOR DE LA MATRIZ ANTERIOR 0,22

NUEVO GRUPO [(K-L),(G-J)]

MATRIZ DE DISTANCIAS MAXIMAS:

	[(A-B),(E-F)]	[(H-I) M]	[(K-L),(G-J)]	C-D
[(A-B),(E-F)]	0	0,82	0,48	0,23
[(H-I), M]		0	0,54	1
[(K-L),(G-J)]			0	0,60
C-D				0

EL MENOR ELEMENTO DE LA MATRIZ ANTERIOR ES 0,23

NUEVO GRUPO {[(A-B),(E-F)],(C-D)}

Y LA CORRESPONDIENTE MATRIZ DE DISTANCIAS:

	{[(A-B),(E-F)],(C-D)}	[(H-I),M]	[(K-L),(G-J)]
{[(A-B),(E-F)],(C-D)}	0	1	0,60
[(H-I),M]		0	0,54
[(K-L),(G-J)]			0

EL MENOR ELEMENTO DE LA MATRIZ ANTERIOR ES 0,54

NUEVO GRUPO {[(H-I),M],[(K-L),(G-J)]}

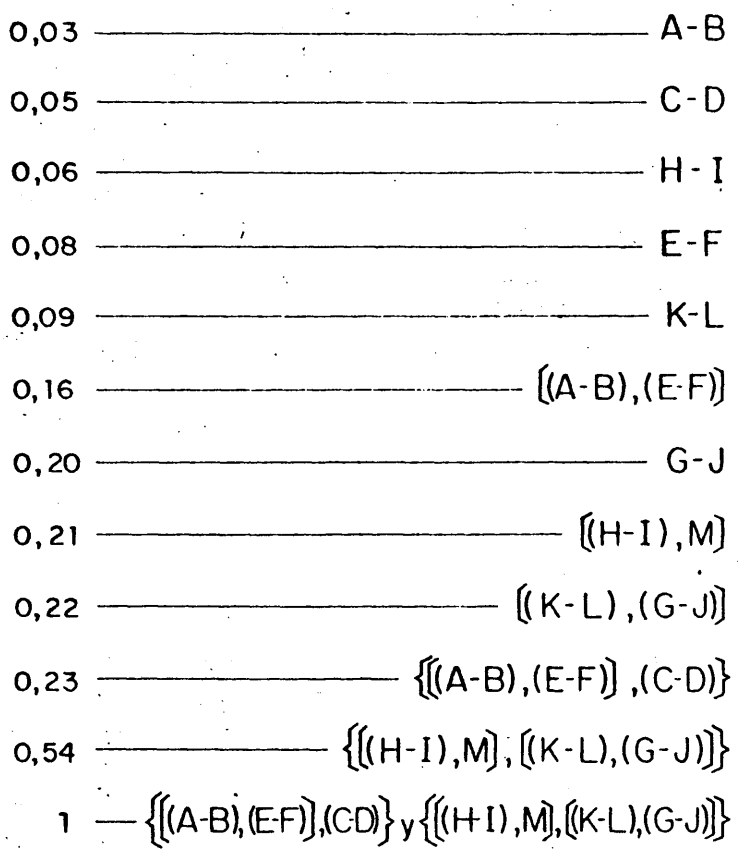
LA CORRESPONDIENTE MATRIZ DE DISTANCIAS MAXIMAS ES:

	{[(A-B),(E-F)],(C-D)}	{[(H-I),M],[(K-L),(G-J)]}
{[(A-B),(E-F)],(C-D)}	0	1
{[(H-I),M],[(K-L),(G-J)]}		0

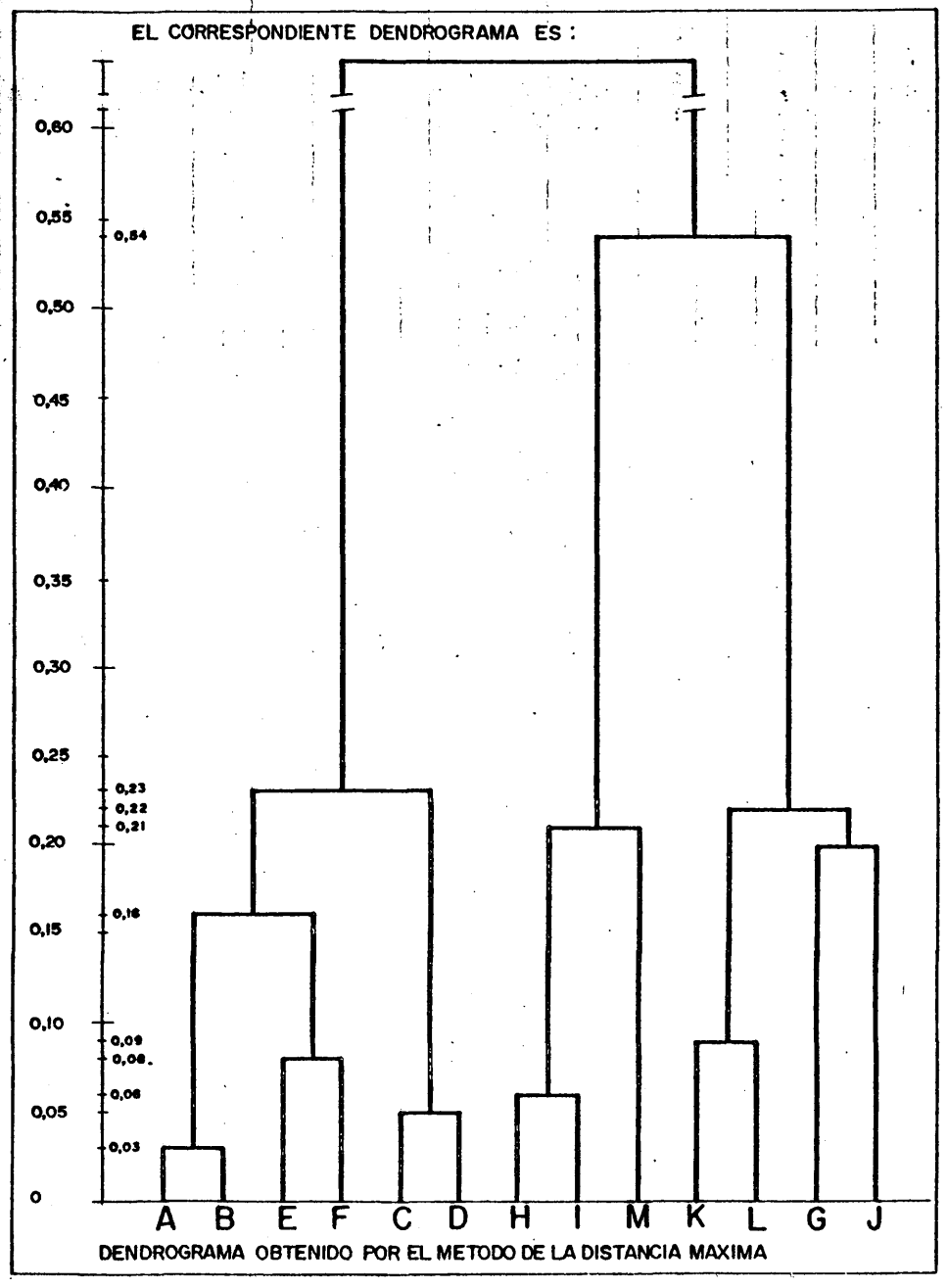
LUEGO LOS GRUPOS {[(A-B),(E-F)],(C-D)} Y {[(H-I)M],[(K-L)(G-J)

SE UNEN AL NIVEL 1

RESUMEN


0,03 A-B
0,05 C-D
0,06 H-I
0,08 E-F
0,09 K-L
0,16 [(A-B),(E-F)]
0,20 G-J
0,21 [(H-I),M]
0,22 [(K-L),(G-J)]
0,23 {[(A-B),(E-F)],(C-D)}
0,54 {[(H-I),M],[(K-L),(G-J)]}
1 {[(A-B),(E-F)],(CD)} y {[(H-I),M],[(K-L),(G-J)]}

EL CORRESPONDIENTE DENDROGRAMA ES :



DENDROGRAMA OBTENIDO POR EL METODO DE LA DISTANCIA MAXIMA

METODO DE LA MEDIA

ELEMENTO MAS PEQUENO DE LA MATRIZ ANTERIOR 0,03

LUEGO EL PRIMER GRUPO LO FORMAN LOS ELEMENTOS A Y B

LA MATRIZ CORRESPONDIENTE DE DISTANCIAS MEDIAS ES:

	A-B	C	D	E	F	G	H	I	J	K	L	M
A-B	0	0,16	0,14	0,15	0,125	0,34	0,65	0,76	0,475	0,30	0,26	0,81
C		0	0,05	0,22	0,23	0,45	0,76	0,86	0,60	0,41	0,40	0,94
D			0	0,18	0,19	0,40	0,75	0,88	0,57	0,37	0,37	1
E				0	0,08	0,25	0,59	0,66	0,41	0,20	0,20	0,71
F					0	0,23	0,57	0,64	0,38	0,19	0,19	0,71
G						0	0,391	0,49	0,199	0,22	0,18	0,51
H							0	0,06	0,22	0,40	0,38	0,21
I								0	0,29	0,47	0,45	0,21
J									0	0,21	0,22	0,33
K										0	0,09	0,52
L											0	0,54
M												0

ELEMENTO MAS PEQUEÑO DE LA MATRIZ ANTERIOR 0,05

NUEVO GRUPO: C-D

LA NUEVA MATRIZ DE DISTANCIAS MEDIAS ES:

	A-B	C-D	E	F	G	H	I	J	K	L	M
A-B	0	0,15	0,15	0,125	0,34	0,65	0,76	0,475	0,30	0,26	0,81
C-D		0	0,20	0,21	0,425	0,755	0,87	0,585	0,39	0,385	0,97
E			0	0,08	0,25	0,59	0,66	0,41	0,20	0,20	0,71
F				0	0,23	0,57	0,64	0,38	0,19	0,19	0,71
G					0	0,391	0,49	0,199	0,22	0,18	0,51
H						0	0,06	0,22	0,40	0,38	0,21
I							0	0,29	0,47	0,45	0,21
J								0	0,21	0,22	0,33
K									0	0,09	0,52
L										0	0,54
M											0

EL MENOR ELEMENTO DE LA MATRIZ ANTERIOR ES 0,06

NUEVO GRUPO H-I

Y LA CORRESPONDIENTE MATRIZ DE DISTANCIAS MEDIAS ES:

	A-B	C-D	H-I	E	F	G	J	K	L	M
A-B	0	0,15	0,705	0,15	0,125	0,34	0,475	0,30	0,26	0,81
C-D		0	0,8125	0,20	0,21	0,425	0,585	0,39	0,385	0,97
H-I			0	0,625	0,605	0,4405	0,255	0,435	0,415	0,21
E				0	0,08	0,25	0,41	0,20	0,20	0,71
F					0	0,23	0,38	0,19	0,19	0,71
G						0	0,199	0,22	0,18	0,51
J							0	0,22	0,18	0,51
K								0	0,09	0,52
L									0	0,54
M										0

EL MENOR ELEMENTO DE LA MATRIZ ANTERIOR ES. 0,08

NUEVO GRUPO E-F

Y LA CORRESPONDIENTE MATRIZ DE DISTANCIAS MEDIAS ES:

	A-B	C-D	H-I	E-F	G	J	K	L	M
A-B	0	0,15	0,705	0,1375	0,34	,4	0,30	0,26	0,81
C-D		0	0,8125	0,205	,425		0,39	0,38	0,97
H-I			0	0,615		0,255	,435	,415	0,21
E-F				0	0,24	0,385	,195	0,195	0,71
G					0	199	0,22	0,18	0,51
J						0	0,22	0,18	0,51
K							0	0,09	0,52
L								0	0,54
M									0

EL MENOR ELEMENTO DE LA MATRIZ ANTERIOR ES 0,09

NUEVO GRUPO K-L

Y LA CORRESPONDIENTE MATRIZ DE DISTANCIAS MEDIAS ES:

	A-B	C-D	H-I	E-F	K-L	G	J	M
A-B	0	0,15	0,705	0,1375	0,28	0,34	0,475	0,81
C-D		0	0,8125	0,205	0,3875	0,425	0,585	0,97
H-I			0	0,615	0,425	0,4406	0,255	0,21
E-F				0	0,195	0,24	0,395	0,71
K-L					0	0,20	0,20	0,53
G						0	0,199	0,51
J							0	0,51
M								0

EN LA MATRIZ ANTERIOR EL ELEMENTO MENOR ES 0,1375

LUEGO EL NUEVO GRUPO ES : [(A-B),(E-F)]

LA CORRESPONDIENTE MATRIZ DE DISTANCIAS ES:

	[(A-B),(E-F)]	C-D	H-I	K-L	G	J	M
[(A-B),(E-F)]	0	0,1775	0,66	0,2375	0,29	0,435	0,76
C-D		0	0,8125	0,3875	0,425	0,585	0,97
H-I			0	0,425	0,4405	0,255	0,21
K-L				0	0,20	0,20	0,53
G					0	0,199	0,51
J						0	0,51
M							0

MENOR ELEMENTO DE LA MATRIZ ANTERIOR ES 0,1775

LUEGO EL NUEVO GRUPO ES : {[(A-B),(E-F)],(C-D)}

Y LA NUEVA MATRIZ DE DISTANCIAS ES:

	{[(A-B),(E-F)],(C-D)}	H-I	K-L	G	J	M
{[(A-B),(E-F)],(C-D)}	0	0,73625	0,3125	0,3575	0,51	0,865
H-I		0	0,425	0,4405	0,255	0,21
K-L			0	0,20	0,20	0,53
G				0	0,199	0,51
J					0	0,51
M						0

ELEMENTO MAS PEQUEÑO DE LA MATRIZ ANTERIOR ES: 0,199

Y EL NUEVO GRUPO ES G-J ; LA NUEVA MATRIZ ES:

	{[(A-B),(E-F)],(C-D)}	H-I	K-L	G-J	M
{[(A-B),(E-F)],(C-D)}	0	0,73625	0,3125	0,43375	0,865
H-I		0	0,425	0,34775	0,21
K-L			0	0,20	0,53
J				0	0,51
M					0

EL ELEMENTO MAS PEQUENO DE LA MATRIZ ANTERIOR ES 0,20

LUEGO EL NUEVO GRUPO A FORMAR ES [(K-L),(G-J)]

Y LA NUEVA MATRID DE DISTANCIAS MEDIAS ES:

	{[(A-B),(E-F)],(C-D)}	[(K-L),(G-J)]	H-I	M
{[(A-B),(E-F)],(C-D)}	0	0,373125	0,73625	0,865
[(K-L),(G-J)]		0	0,386375	0,52
H-I			0	0,21
M				0

MENOR ELEMENTO 0,21

NUEVO GRUPO (H-I), M

MATRIZ DE DISTANCIAS:

	{[(A-B),(E-F)],(C-D)}	[(K-L),(G-J)]	[(H-I),M]
{[(A-B),(E-F)],(C-D)}	0	0,373125	0,800625
[(K-L),(G-J)]		0	0,4531875
[(H-I),M]			0

MENOR ELEMENTO 0,373125

NUEVO GRUPO |{[(A-B),(E-F)],(C-D)},[(K-L),(G-J)]|

MATRIZ DE DISTANCIAS:

	|{[(A-B),(E-F)],(C-D)},[(K-L),(G-J)]|	[(H-I),M]
|{[(A-B),(E-F)],(C-D)},[(K-L),(G-J)]|	0	0,62690625
[(H-I),M]			0

RESUMEN

0,03 —— A-B

0,05 —— C-D

0,06 —— H-I

0,08 —— E-F

0,09 —— K-L

0,1375 —— [(A-B),(E-F)]

0,1775 —— {[(A-B),(E-F)],(C-D)}

0,199 —— G-J

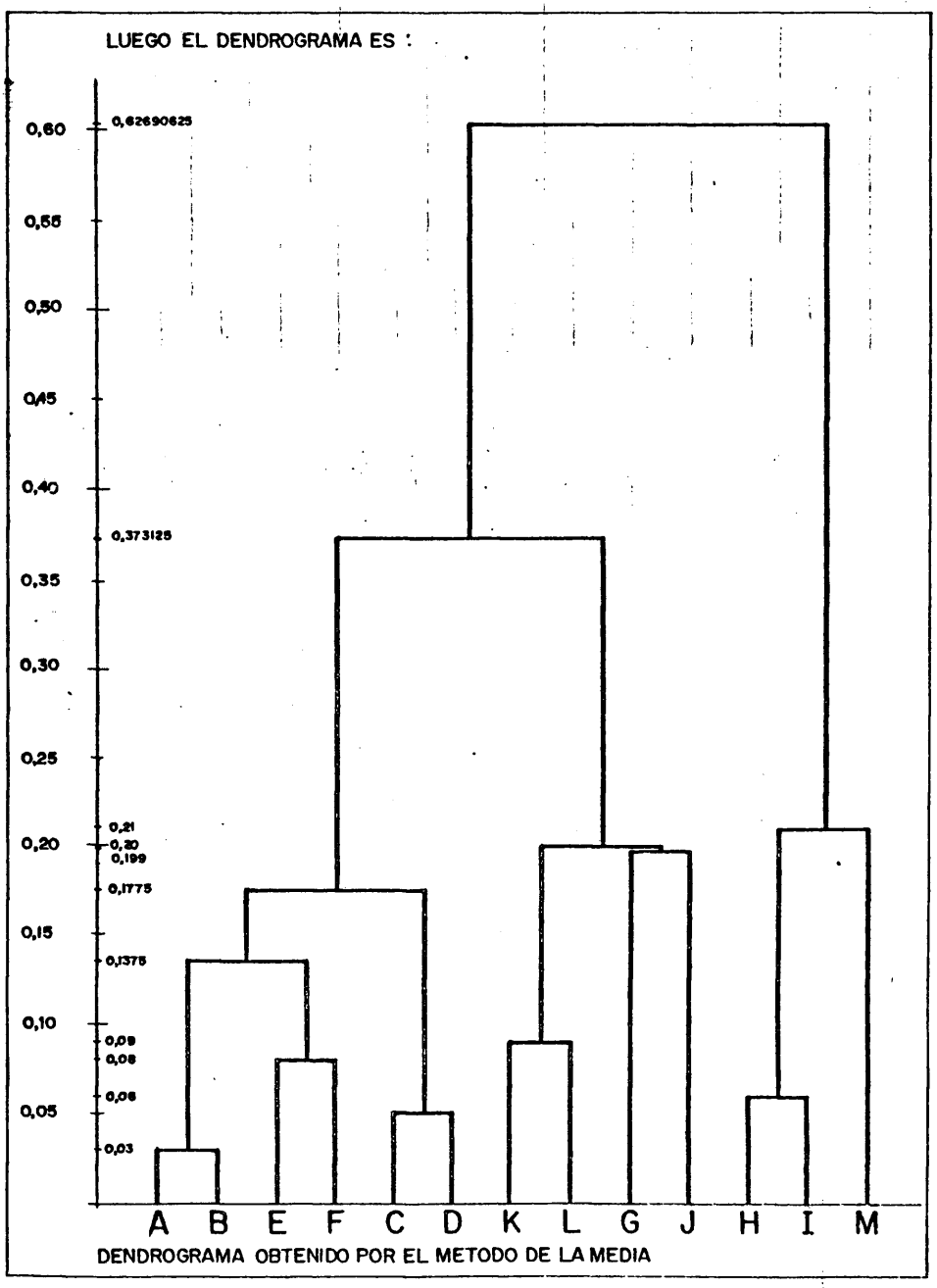
0,20 —— [(K-L),(G-J)]

0,21 —— [(H-I),M]

0,373125 —— |{[(A-B),(E-F)],(C-D)},[(K-L),(G-J)]|

0,62690625 —— |{[(A-B),(E-F)],(C-D)},[(K-L),(G-J)]| con [(H-I),M]

LUEGO EL DENDROGRAMA ES :



DENDROGRAMA OBTENIDO POR EL METODO DE LA MEDIA

METODO DE LA DISTANCIA δ

EL MENOR ELEMENTO DE LA MATRIZ ES 0,03

LUEGO LOS OTUS A Y B FORMAN EL PRIMER GRUPO = A-B

LA MATRIZ DE DISTANCIAS ES:

	A-B	C	D	E	F	G	H	I	J	K	L	M
A-B	0	0,16	0,14	0,15	0,125	0,34	0,65	0,76	0,475	0,30	0,26	0,81
C		0	0,05	0,22	0,23	0,45	0,76	0,86	0,60	0,41	0,40	0,94
D			0	0,18	0,19	0,40	0,75	0,88	0,57	0,37	0,37	1
E				0	0,08	0,25	0,59	0,66	0,41	0,20	0,20	0,71
F					0	0,23	0,57	0,64	0,38	0,19	0,19	0,71
G						0	0,391	0,49	0,199	0,22	0,18	0,51
H							0	0,06	0,22	0,40	0,38	0,21
I								0	0,29	0,47	0,45	0,21
J									0	0,21	0,22	0,33
K										0	0,09	0,52
L											0	0,54
M												0

EL ELEMENTO MAS PEQUEÑO DE LA MATRIZ ANTERIOR ES 0,05

LUEGO EL NUEVO GRUPO ES C-D

Y LA NUEVA MATRIZ ES:

	A-B	C-D	E	F	G	H	I	J	K	L	M
A-B	0	0,10	0,15	0,125	0,34	0,65	0,76	0,475	0,30	0,26	0,81
C-D		0	0,20	0,21	0,425	0,755	0,87	0,585	0,39	0,385	0,97
E			0	0,08	0,25	0,59	0,66	0,41	0,20	0,20	0,71
F				0	0,23	0,57	0,64	0,38	0,19	0,19	0,71
G					0	0,391	0,49	0,20	0,22	0,18	0,51
H						0	0,06	0,199	0,40	0,38	0,21
I							0	0,29	0,47	0,45	0,21
J								0	0,21	0,22	0,33
K									0	0,09	0,52
L										0	0,54
M											0

MENOR ELEMENTO MATRIZ ANTERIOR ES 0,06

NUEVO GRUPO H-I

NUEVA MATRIZ DE DISTANCIAS δ

	A-B	C-D	H-I	E	F	G	J	K	L	M
A-B	0	0,10	0,715	0,15	0,125	0,34	0,475	0,30	0,26	0,81
C-D		0	0,815	0,20	0,21	0,425	0,585	0,39	0,385	0,97
H-I			0	0,625	0,605	0,4405	0,255	0,435	0,415	0,21
E				0	0,08	0,25	0,41	0,20	0,20	0,71
F					0	0,23	0,38	0,19	0,19	0,71
G						0	0,199	0,22	0,18	0,51
J							0	0,22	0,18	0,51
K								0	0,09	0,52
L									0	0,54
M										0

MENOR ELEMENTO MATRIZ ANTERIOR 0,08

NUEVO GRUPO E-F

NUEVA MATRIZ DE DISTANCIAS δ

	A-B	C-D	H-I	E-F	G	J	K	L	M
A-B	0	0,10	0,715	0,14	0,34	0,475	0,30	0,26	0,81
C-D		0	0,815	0,205	0,425	0,585	0,39	0,385	0,97
H-I			0	0,615	0,4405	0,255	0,435	0,415	0,21
E-F				0	0,24	0,395	0,195	0,195	0,71
G					0	0,199	0,22	0,18	0,51
J						0	0,22	0,18	0,51
K							0	0,09	0,52
L								0	0,54
M									0

MENOR ELEMENTO MATRIZ ANTERIOR ES 0,09

NUEVO GRUPO K-L

NUEVA MATRIZ DE DISTANCIAS δ

	A-B	C-D	H-I	E-F	K-L	G	J	M
A-B	0	0,10	0,715	0,14	0,285	0,34	0,475	0,81
C-D		0	0,815	0,205	0,39	0,425	0,585	0,97
H-I			0	0,615	0,425	0,440	0,255	0,21
E-F				0	0,195	0,24	0,395	0,71
K-L					0	0,20	0,20	0,53
G						0	0,199	0,51
J							0	0,51
M								0

MENOR ELEMENTO MATRIZ ANTERIOR 0,10

NUEVO GRUPO A-B-C-D Y LA NUEVA MATRIZ ES:

	A-B-C-D	H-I	E-F	K-L	G	J	M
A-B-C-D	0	0,76	0,175	0,335		,53	
H-I		0	0,615	0,425			0,21
E-F			0	0,195	0,24		0,71
K-L				0	0,20	0,20	0,53
G					0	199	0,51
J						0	0,51
M							0

MENOR ELEMENTO MATRIZ ANTERIOR 0,175

NUEVO GRUPO A-B-C-D-E-F

MATRIZ :

	A-B-C-D-E-F	H-I	K-L	G	J	M
A-B-C-D-E-F	0	0,725	0,3	,34	0,49	
H-I		0	0,425		0,2	0,21
K-L			0	0,2	0,20	0,53
G				0	1	,51
J					0	0,51
M						0

MENOR ELEMENTO MATRIZ ANTERIOR 0,199

NUEVO GRUPO G-J

MATRIZ:

	A-B-C-D-E-F	H-I	K-L	G-J	M
A-B-C-D-E-F	0	0,725	0,3	0,415	0,855
H-I		0	0,425	0,345	0,21
K-L			0	0,20	0,53
G-J				0	0,51
M					0

MENOR ELEMENTO MATRIZ ANTERIOR 0,20

NUEVO GRUPO K-L-G-J

MATRIZ:

	A-B-C-D-E-F	K-L-G-J	H-I	M
A-B-C-D-E-F	0	0,395	0,725	0,855
K-L-G-J		0	0,345	0,435
H-I			0	0,21
M				0

MENOR ELEMENTO MATRIZ ANTERIOR 0,21

NUEVO GRUPO H-I-M

MATRIZ:

	A-B-C-D-E-F	K-L-G-J	H-I-M
A-B-C-D-E-F	0	0,395	0,785
K-L-G-J		0	0,38
H-I-M			0

MENOR ELEMENTO MATRIZ ANTERIOR ES 0,38

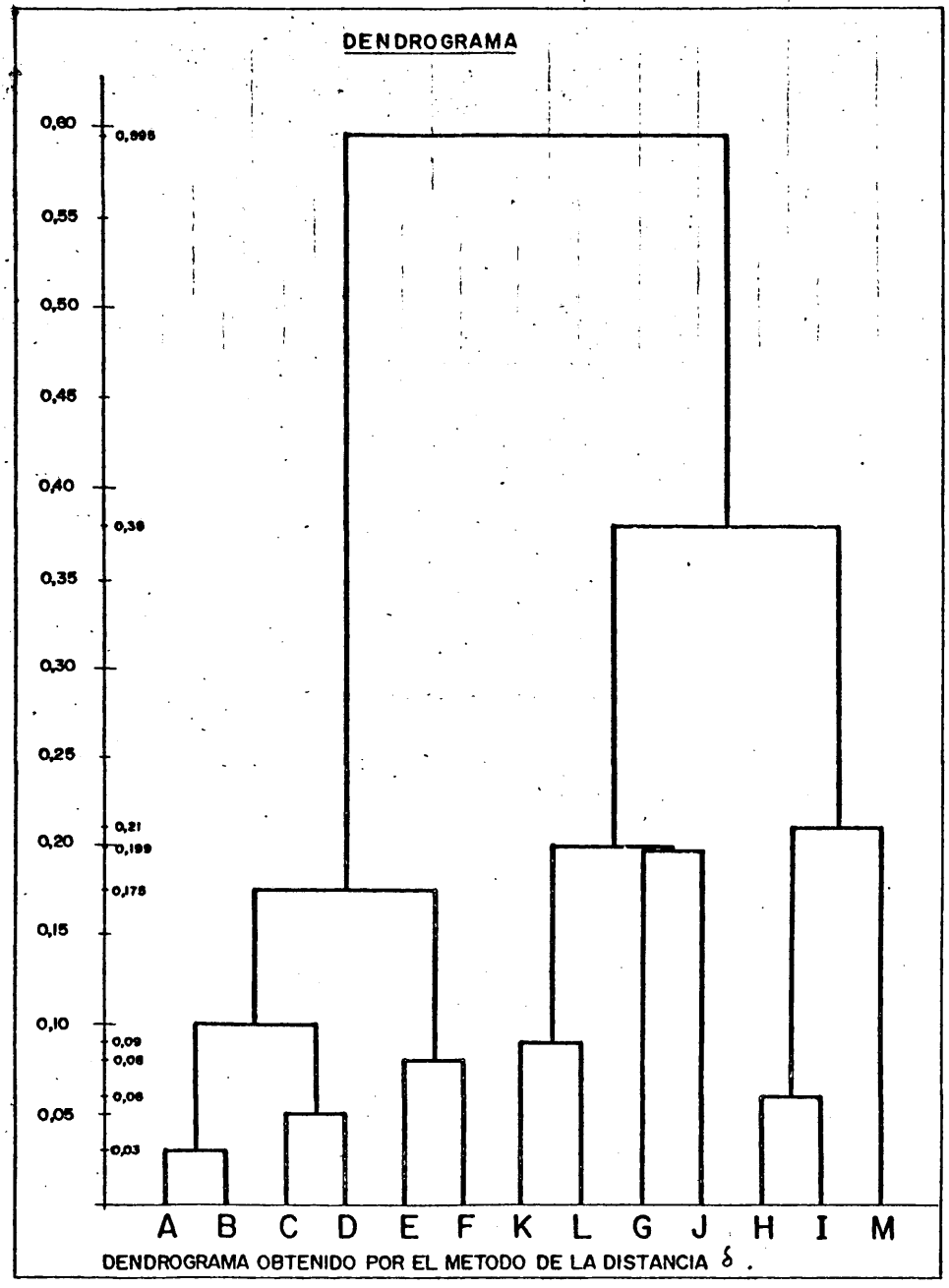
NUEVO GRUPO K-L-G-J-H-I-J

MATRIZ:

	A-B-C-D-E-F	K-L-G-J-H-I-M
A-B-C-D-E-F	0	0,595
K-L-G-J-H-I-M		0

RESUMEN

0,03	A-B
0,05	C-D
0,06	H-I
0,08	E-F
0,09	K-L
0,10	(A-B;C-D)
0,175	[(A-B,C-D),E-F]
0,199	G-J
0,20	(K-L,G-J)
0,21	(H-I,M)
0,38	[(K-L-G-J)] [(H-I),M]
0,595	(A-B-E-D-E-F) (K-L-G-J-H-I,M)



DENDROGRAMA OBTENIDO POR EL METODO DE LA DISTANCIA δ .

Una vez obtenidos los dendrogramas anteriores, es decir, terminado el proceso de la conglomeración hemos de trabajar en -- las direcciones siguientes:

a.- Interpretación de los resultados.

b.- Estudiar la relación existente entre los datos originales y los resultados de la conglomeración.

c.- Comparación de los dendrogramas.

Los apartados b y c constituyen la segunda parte de este capítulo y en lo que respecta al apartado a entendemos que debe - ser el científico especializado, en este caso el Dr. Talens quien comenta los resultados obtenidos, por lo que transcribimos literalmente el informe que nos ha facilitado, y que se refiere al dendrograma obtenido en primer lugar, por el método de la distancia mínima ya que, como veremos en la parte siguiente es el que nos ha proporcionado el mayor coeficiente de correlación al comparar sus resultados con los datos originales.

"El dendrograma obtenido es esclarecedor en general, respecto a las relaciones de las formas, pero, también, indica posibles desviaciones, o una "mala clasificación" de alguna de ellas.

Así: Los ejemplares A y B con índice de 0,03 son claramente de la misma especie: Neuropteris heterophyla.

Lo mismo pasa para los E y F con 0,08 que son: N. -- scheuzeri, y C y D con 0,05 que son N. temuifolia.

Ocurre lo mismo para el K y L que son Linopteris neuropteroides con 0,09, género de forma que está perfectamente dentro del grupo y debe considerarse como "especie" dentro del parataxón. Sin embargo, su separación o nivel respecto al grueso del grupo - que forma el género es notablemente grande con un índice 0,19.

Los ejemplares del H e I son, también, de modo claro, de la misma especie: Neuropteris linguaefolia que incluso parece un grupo altamente aparte del parataxón.

El ejemplar J que, clasificado según costumbre, pertenece también al N. linguaefolia presenta un índice de 0,22 respecto a los que constituyen el grueso de la especie.

Lo mismo ocurre con los G y M.

Este hecho puede proceder de errores en la toma de medidas, a malas interpretaciones en las fotografías o, incluso, a -- una mala clasificación clásica.

En este aspecto, es de destacar que pese a toda la objetividad deseada por los sistemáticos y las precauciones (holotipos, dibujos, fotografías y descripciones publicadas) tomadas, muchas - veces para cada autor (leasé paleontólógo) cada una de las especies queda como una representación ideal, casi un "sentimiento".

Es, precisamente, para obviar este factor subjetivo en - la clasificación de fósiles, por lo que se hace necesaria la aplicación de los métodos de la taxonomía numérica!

COMPARACION DE LOS RESULTADOS

Como dijimos al final del apartado A, de este capítulo, hemos de resolver aún las dos cuestiones siguientes:

1.- ¿Qué relación existe entre los datos originales y las clasificaciones proporcionadas por los dendrogramas?

2.- Comparación de los dendrogramas anteriores.

La cuestión que planteamos en primer lugar la hemos resuelto mediante la utilización del coeficiente de correlación de SOKAL.

En primer lugar hemos obtenido, a partir de cada dendrograma la matriz de valores cofenéticos y en segundo lugar hemos calculado, en cada uno de los cuatro casos el coeficiente de correlación cofenético entre esta matriz y la matriz de los datos originales.

Los resultados, en cada caso, han sido los siguientes:

	A	B	C	D	E	F	G	H	I	J	K	L	M
	3	3,2	2,3	2,6	3,7	3,9	4,9	6,6	6,9	5,7	4,8	4,6	7,8
	2,1	2,2	1,3	1,2	1,3	1,7	1,4	3,3	3,7	2,4	1,6	2,2	2,6
	1,2	1,4	1,4	1,2	1,7	1,2	0,7	2	2,1	1,1	1,4	1,8	1,4
	2,3	2,3	1,8	2	2,8	2,7	4,1	5,8	6,1	5	3,9	3,8	6,6
A 3-2,1-1,2-2,3	0	0,03	0,15	0,13	0,16	0,13	0,34	0,66	0,73	0,48	0,31	0,26	0,82
B 3,2-2,2-1,4-2,3	0,03	0	0,17	0,15	0,14	0,12	0,34	0,64	0,79	0,47	0,29	0,26	0,80
C 2,3-1,3-1,4-1,8	0,15	0,17	0	0,05	0,22	0,23	0,45	0,76	0,86	0,60	0,41	0,4	0,94
D 2,6-1,2-1,2-2	0,13	0,15	0,05	0	0,18	0,19	0,40	0,75	0,88	0,57	0,37	0,37	1
E 3,7-1,3-1,7-2,8	0,16	0,14	0,22	0,18	0	0,08	0,25	0,59	0,66	0,41	0,20	0,20	0,71
F 3,9-1,7-1,2-2,7	0,13	0,12	0,23	0,19	0,08	0	0,23	0,57	0,64	0,38	0,19	0,19	0,71
G 4,9-1,4-0,7-4,1	0,34	0,34	0,45	0,40	0,25	0,23	0	0,391	0,49	0,20	0,22	0,18	0,51
H 6,6-3,3-2-5,8	0,66	0,64	0,76	0,75	0,59	0,57	0,391	0	0,06	0,22	0,40	0,38	0,21
I 6,9-3,7-2,1-6,1	0,73	0,79	0,86	0,88	0,66	0,64	0,49	0,06	0	0,29	0,47	0,45	0,21
J 5,7-2,4-1,1-5	0,48	0,47	0,60	0,57	0,41	0,38	0,20	0,22	0,29	0	0,21	0,22	0,33
K 4,8-1,6-1,4-3,9	0,31	0,29	0,41	0,37	0,20	0,19	0,22	0,40	0,47	0,21	0	0,09	0,52
L 4,6-2,2-1,8-3,8	0,26	0,26	0,4	0,37	0,20	0,19	0,18	0,38	0,45	0,22	0,09	0	0,54
M 7,8-2,6-1,4-6,6	0,82	0,80	0,94	1	0,71	0,71	0,51	0,21	0,21	0,33	0,52	0,54	0

MATRIZ DE DISTANCIAS EUCLIDEAS OBTENIDAS A PARTIR DE LOS DATOS ORIGINALES

		A	B	C	D	E	F	G	H	I	J	K	L	M
		3	3,2	2,3	2,6	3,7	3,9	4,9	6,6	6,9	5,7	4,8	4,6	7,8
		2,1	2,2	1,3	1,2	1,3	1,7	1,4	3,3	3,7	2,4	1,6	2,2	2,6
		1,2	1,4	1,4	1,2	1,7	1,2	0,7	2	2,1	1,1	1,4	1,8	1,4
		2,3	2,3	1,8	2	2,8	2,7	4,1	5,8	6,1	5	3,9	3,8	6,6
A	3-2,1-1,2-2,3	0	0,03	0,13	0,13	0,12	0,12	0,19	0,22	0,22	0,20	0,19	0,19	0,22
B	3,2-2,2-1,4-2,3	0,03	0	0,13	0,13	0,12	0,12	0,19	0,22	0,22	0,20	0,19	0,19	0,22
C	2,3-1,3-1,4-1,8	0,13	0,13	0	0,05	0,13	0,13	0,19	0,22	0,22	0,2	0,19	0,19	0,22
D	2,6-1,2-1,2-2	0,13	0,13	0,05	0	0,13	0,13	0,19	0,22	0,22	0,2	0,19	0,19	0,22
E	3,7-1,3-1,7-2,8	0,12	0,12	0,13	0,13	0	0,08	0,19	0,22	0,22	0,2	0,19	0,19	0,22
F	3,9-1,7-1,2-2,7	0,12	0,12	0,13	0,13	0,08	0	0,19	0,22	0,22	0,2	0,19	0,19	0,22
G	4,9-1,4-0,7-4,1	0,19	0,19	0,19	0,19	0,19	0,19	0	0,22	0,22	0,2	0,18	0,18	0,22
H	6,6-3,3-2-5,8	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0	0,06	0,22	0,22	0,22	0,21
I	6,9-3,7-2,1-6,1	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,06	0	0,22	0,22	0,22	0,21
J	5,7-2,4-1,1-5	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,22	0,22	0	0,20	0,2	0,22
K	4,8-1,6-1,4-3,9	0,19	0,19	0,19	0,19	0,19	0,19	0,18	0,22	0,22	0,20	0	0,09	0,22
L	4,6-2,2-1,8-3,8	0,19	0,19	0,19	0,19	0,19	0,19	0,18	0,22	0,22	0,2	0,09	0	0,22
M	7,8-2,6-1,4-6,6	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,21	0,21	0,22	0,22	0,22	0

MATRIZ DE VALORES COFENETICOS CORRESPONDIENTE AL DENDROGRAMA OBTENIDO POR EL METODO DE LA DISTANCIA MINIMA.

		A	B	C	D	E	F	G	H	I	J	K	L	M
		3	3,2	2,3	2,6	3,7	3,9	4,9	6,6	6,9	5,7	4,8	4,6	7,8
		2,1	2,2	1,3	1,2	1,3	1,7	1,4	3,3	3,7	2,4	1,6	2,2	2,6
		1,2	1,4	1,4	1,2	1,7	1,2	0,7	2	2,1	1,1	1,4	1,8	1,4
		2,3	2,3	1,8	2	2,8	2,7	4,1	5,8	6,1	5	3,9	3,8	6,6
A	3-2,1-1,2-2,3	0	0,03	0,23	0,23	0,16	0,16	1	1	1	1	1	1	1
B	3,2-2,2-1,4-2,3	0,03	0	0,23	0,23	0,16	0,16	1	1	1	1	1	1	1
C	2,3-1,3-1,4-1,8	0,23	0,23	0	0,05	0,23	0,23	1	1	1	1	1	1	1
D	2,6-1,2-1,2-2	,23	0,23	0,	0	0,23	0,23	1	1	1	1	1	1	1
E	3,7-1,3-1,7-2,8	0,16	0,16	,23	0,23	0	0,08	1	1	1	1	1	1	1
F	3,9-1,7-1,2-2,7	0,16	0,16	0,23	0,23	0,08	0	1	1	1	1	1	1	1
G	4,9-1,4-0,7-4,1	1	1	1	1	1	1	0	0,54	0,54	0,20	0,22	,22	0,54
H	6,6-3,3-2-5,8	1	1	1	1	1	1	0,54	0	0,06	0,54	0,54	,54	0,21
I	6,9-3,7-2,1-6,1	1	1	1	1	1	1	0,54	0,0	0	0,54	0,54	0,54	0,21
J	5,7-2,4-1,1-5	1	1	1	1	1	1	0,20	0,54	0,54	0	0,22	0,22	0,54
K	4,8-1,6-1,4-3,9	1	1	1	1	1	1	0,22	0,54	0,54	0,22	0	0,0	0,54
L	4,6-2,2-1,8-3,8	1	1	1	1	1	1	,22	,54	0,54	0,22	0,09	0	0,54
M	7,8-2,6-1,4-6,6	1	1	1	1	1	1	54	0,21	0,21	0,54	0,54	,54	0

MATRIZ DE VALORES COFENETICOS CORRESPONDIENTE AL DENDROGRAMA OBTENIDO POR EL METODO DE LA DISTANCIA MAXIMA

		A	B	C	D	E	F	G	H	I	J	K	L	M
		3	3,2	2,3	2,6	3,7	3,9	4,9	6,6	6,9	5,7	4,8	4,6	7,8
		2,1	2,2	1,3	1,2	1,3	1,7	1,4	3,3	3,7	2,4	1,6	2,2	2,6
		1,2	1,4	1,4	1,2	1,7	1,2	0,7	2	2,1	1,1	1,4	1,8	1,4
		2,3	2,3	1,8	2	2,8	2,7	4,1	5,8	6,1	5	3,9	3,8	6,6
A	3-2,1-1,2-2,3	0	0,03	0,1775	0,1775	0,1375	0,1375	0,3731	0,6270	0,6270	0,3731	0,3731	0,3731	0,6270
B	3,2-2,2-1,4-2,3	0,03	0	0,1775	0,1775	0,1375	0,1375	0,3731	0,6270	0,6270	0,3731	0,3731	0,3731	0,6270
C	2,3-1,3-1,4-1,8	0,1775	0,1775	0	0,06	0,1775	0,1775	0,3731	0,6270	0,6270	0,3731	0,3731	0,3731	0,6270
D	2,6-1,2-1,2-2	0,1775	0,1775	0,06	0	0,1775	0,1775	0,3731	0,6270	0,6270	0,3731	0,3731	0,3731	0,6270
E	3,7-1,3-1,7-2,8	0,1375	0,1375	0,1775	0,1775	0	0,08	0,3731	0,6270	0,6270	0,3731	0,3731	0,3731	0,6270
F	3,9-1,7-1,2-2,7	0,1375	0,1375	0,1775	0,1775	0,08	0	0,3731	0,6270	0,6270	0,3731	0,3731	0,3731	0,6270
G	4,9-1,4-0,7-4,1	0,3731	0,3731	0,3731	0,3731	0,3731	0,3731	0	0,6270	0,6270	0,199	0,20	0,20	0,6270
H	6,6-3,3-2-5,8	0,6270	0,6270	0,6270	0,6270	0,6270	0,6270	0,6270	0	0,06	0,6270	0,6270	0,6270	0,21
I	6,9-3,7-2,1-6,1	0,6270	0,6270	0,6270	0,6270	0,6270	0,6270	0,6270	0,06	0	0,6270	0,6270	0,6270	0,21
J	5,7-2,4-1,1-5	0,3731	0,3731	0,3731	0,3731	0,3731	0,3731	0,199	0,6270	0,6270	0	0,20	0,20	0,6270
K	4,8-1,6-1,4-3,9	0,3731	0,3731	0,3731	0,3731	0,3731	0,3731	0,20	0,6270	0,6270	0,20	0	0,09	0,6270
L	4,6-2,2-1,8-3,8	0,3731	0,3731	0,3731	0,3731	0,3731	0,3731	0,20	0,6270	0,6270	0,20	0,09	0	0,6270
M	7,8-2,6-1,4-6,6	0,6270	0,6270	0,6270	0,6270	0,6270	0,6270	0,6270	0,21	0,21	0,6270	0,6270	0,6270	0

MATRIZ DE VALORES COFENETICOS CORRESPONDIENTE AL DENDROGRAMA OBTENIDO POR EL METODO DE LA MEDIA

		A	B	C	D	E	F	G	H	I	J	K	L	M
		3	3,2	2,3	2,6	3,7	3,9	4,9	6,6	6,9	5,7	4,8	4,6	7,8
		2,1	2,2	1,3	1,2	1,3	1,7	1,4	3,3	3,7	2,4	1,6	2,2	2,6
		1,2	1,4	1,4	1,2	1,7	1,2	0,7	2	2,1	1,1	1,4	1,8	1,4
		2,3	2,3	1,8	2	2,8	2,7	4,1	5,8	6,1	5	3,9	3,8	6,6
A	3-2,1-1,2-2,3	0	0,03	0,1	0,1	0,175	0,175	0,595	0,595	0,5	,59	0,595	,595	,595
B	3,2-2,2-1,4-2,3	,03	0	0,1	0,1	0,175	0,175	0,595	0,595	0,595	,595	0,595	0,595	0,595
C	2,3-1,3-1,4-1,8	0,1	0,1	0	0,05	0,175	0,175	,595	0,595	0,5	0,5	0,5	0,595	,595
D	2,6-1,2-1,2-2	0,1	0,1	0,05	0	0,175	0,175	0,595	0,595	0,595	,5	,595	,595	0,5
E	3,7-1,3-1,7-2,8	0,175	0,175	0,175	0,175	0	0,08	0,595	,595	9	,595	0,5	0,595	0,595
F	3,9-1,7-1,2-2,7	0,175	,175	0,175	,175	0,08	0	0,595	,595	95		0,595	0,595	5
G	4,9-1,4-0,7-4,1	,595	0,595	0,5	,595	0,595	,595	0	0,38	0,38	,199	0,20	0,20	0,38
H	6,6-3,3-2-5,8	95	0,595	0,59	0,595	0,595	,595	0,38	0	0,06	0,38	0,38	0,38	0,21
I	6,9-3,7-2,1-6,1	,5	0,595	0,5	0,59	0,59		0,38	0,06	0	0,38	0,38	0,38	0,21
J	5,7-2,4-1,1-5	0	0,595	0,5	0,5	0,59	0	0,199	0,38	0,38	0	0,20	0,20	0,38
K	4,8-1,6-1,4-3,9	,5	0,595	0	0,595	0,59	0,595	0,20	0,38	0,38	0,20	0	0,09	0,38
L	4,6-2,2-1,8-3,8	0,595	0,595	0,59	0,5	0,5	0,595	0,20	0,38	0,38	0,20	0,09	0	0,38
M	7,8-2,6-1,4-6,6		0,595	,59	0,5	0	0,59	0,38	0,21	0,21	0,38	0,38	0,38	0

MATRIZ DE VALORES COFENETICOS CORRESPONDIENTE AL DENDROGRAMA OBTENIDO POR EL METODO DE LA DISTANCIA δ

Al comparar la matriz de distancias, d, euclideas con las matrices de valores cofenéticos obtenidas, respectivamente, por los métodos de la distancia mínima, m, distancia máxima, n, método de la media, m , y nuestro método δ hemos obtenido los siguientes coeficientes de correlación:

$r_{dm} = 0,724$

$r_{dM} = -0,391$

$r_{dm'} = 0,601$

$r_{d\delta} = 0,706$

por lo que podemos hacer los siguientes comentarios:

a.- El método que proporciona mayor correlación entre sus resultados y los datos originales es el de la distancia mínima.

b.- Nuestro método, δ, obtiene, también buenos resultados clasificatorios puesto que el coeficiente cofenético obtenido, 0,706, difiere muy poco del obtenido por el método de la distancia mínima.

c.- El método que mayor distorsión produce a la hora de clasificar es el método de la distancia máxima; el coeficiente de correlación obtenido, muy bajo, --0,391 nos indica que la covarianza es negativa.

d.- El método de la media, como su nombre indica, y como es lógico esperar por su estructura nos proporciona una clasificación de los elementos que es intermedia entre los métodos de la distancia mínima y de la distancia máxima.

e.- La inspección ocular de los dendrogramas anteriores corrobora los comentarios anteriores, tanto en la conglomeración a niveles inferiores, como a los más altos niveles.

CAPITULO VI

ESTUDIO COMPARATIVO DE METRICAS

INTRODUCCION

Al abordar este problema de la comparación de métricas surge, de una forma lógica, la siguiente cuestión ¿existe alguna experiencia satisfactoria en este campo? y en caso afirmativo; - qué procedimiento se siguió para resolver tal cuestión?; ¿cuales fueron los fundamentos teóricos de tal proceso?; ¿qué resultados se obtuvieron?; ¿qué se deduce de los mismos?

Como las cuestiones anteriores presentes y de un modo ordenado, exponemos esquemáticamente lo siguiente:

1.- Las palabras clave del enunciado del problema son tres: comparación, métrica y taxonomía.

2.- El punto anterior nos conduce, de forma lógica al siguiente proceso escalonado de razonamiento:

   a.- Es necesaria una comparación, y posteriormente una evaluación de los resultados de los métodos jerárquicos del análisis de conglomerados puesto que son muchos los resultados obtenidos hasta el momento presente y se impone su clasificación y ordenación.

   b.- También se necesitan conceptos generales de medición, en particular de la similitud y disimilitud entre elementos lo que supone una sistematización teórica y un catálogo de medidas y sus propiedades.

c.- Al ser la taxonomía la ciencia de la clasificación, y una de sus ramas, el análisis de conglomerados, que estudia la formación de clases, necesitamos conocimientos teóricos básicos de tales ciencias.

d.- Finalmente es también necesario un conocimiento de dos métodos de conglomeración, en particular de los jerárquicos y de su comparación y evaluación,

Todos los apartados anteriores constituyen el marco científico en el que se desarrollará nuestro problema y sus distintos aspectos fundamentales han motivado los capítulos expuestos anteriormente.

Podemos afirmar que no conocemos antecedentes bibliográficos en los que se lleve a cabo un estudio comparativo de distintas métricas; como afirma SNEATH y SOKAL (112).

"Aún no tenemos un estudio sistemático acerca de la comparación de los efectos de los diferentes métodos de conglomeración ya estudiados. Por esta razón la discusión que sigue debe basarse - en consideraciones teóricas cuya significación debe ser validada - empiricamente".

La realidad es que, aunque la frase anterior se escribió en 1963 seguimos en la misma situación; por tal razón hemos intentado, con las líneas que siguen, aportar un procedimiento de resolución y unos resultados al problema planteado de la compara-

ción de métricas y ante los sorprendentes resultados obtenidos en una primera etapa, al estudiar elementos del grupo de los no metales del sistema periódico, quisimos corroborar y contrastarlos con un nuevo ejemplo totalmente diferente del anterior y elegimos los ejemplares, ya estudiados, de NEUROPTERIS.

B.- PLANTEAMIENTO DEL PROBLEMA

DEFINICION: Sea E un conjunto no vacío.
Se dice que una aplicación d de E x E en R (d: E x E⟶R) es una DISTANCIA sobre E si verifica las siguientes propiedades:

a) $d(x,y) = 0$ es equivalente a $x = y$, $d(x,y) \geq 0$
b) $d(x,y) = d(y,x)$
c) $d(x,y) \leq d(x,z)+d(z,y)$, cualesquiera que sean $x,y,z \in E$.

DEFINICION: Al par (E,d) se le llama ESPACIO METRICO.
Sea M(X) el conjunto de dos métricas (distancias) definidas sobre el conjunto X y sea $M_1 = \{d_1,d_2,d_3,d_4\}$ $M \subset M$ un conjunto cuyos elementos son los siguientes:

d_1 es la distancia euclidea, d_2 la distancia de manzanas (City-block), d_3 la distancia de Chebychev y d_4 la distancia de Canberra, definidas en el capítulo II.

Puesto que deseamos comparar las métricas anteriores, definimos en M_1 la siguiente métrica:

$$\delta(d_i,d_j) = \max\left\{\left|d_i(x,y) - d_j(x,y)\right| : x,y \in X \right.$$
$$\left. i,j = 1,2,3,4, \quad i \neq j \right\}$$

Puesto que vamos a hacer una comparación de métricas, calculadas de diferentes formas, dicha comparación ha de ser independiente de los factores de escala en las métricas por lo que habrá que normalizarlas y, a continuación, utilizar la métrica adecuada para compararlas.

$$\delta^{*}(d_i, d_j) = \delta\left[d_i / So(d_i) \ , \ d_j / So(d_j) \right] =$$

$$= \max \left\{ \left| \frac{d_i(x,y)}{So(d_i)} - \frac{d_j(x,y)}{So(d_j)} \right| \ x,y \in X \right.$$

$$\left. i,j = 1,2,3,4, \quad i \neq j \right\}$$

siendo $So(d_i) = \delta(d_i, \overline{0})$ y siendo $\overline{0}$ tal que $\overline{0}(x,y) = 0$, $\forall x,y \in X$

A continuación, definimos los dos conjuntos de OTUS que vamos a utilizar para la comparación de las métricas indicadas:

1. X va a ser el conjunto de los elementos no metales del sistema periódico, es decir,

$$X = \left\{ {}^{6}_{12}C, {}^{7}_{14}N, {}^{8}_{16}O, {}^{9}_{19}F, {}^{15}_{31}P, {}^{16}_{32}S, {}^{17}_{35}Dl, {}^{34}_{79}Se, {}^{35}_{80}Br, {}^{52}_{128}Te, {}^{53}_{127}I, {}^{85}_{210}At \right\}$$

en donde a cada elemento le hemos asociado como caracteres, un subíndice, su peso atómico, y un superíndice, su número atómico, que lo identifican perfectamente.

2. En esta ocasión, el conjunto X va a estar constituído por los trece ejemplares de Neuropteris con los que hemos trabajado en el capítulo V. Cada ejemplar está definido por cuatro caracteres morfológicos.

RESOLUCION DEL PROBLEMA

1. En primer lugar, hemos obtenido las matrices de distancias correspondientes a las distintas métricas aplicadas a los elementos del conjunto X, que ofrecemos en las páginas siguientes:

EUCLIDEA:

$$d_1(\bar{x},\bar{y}) = \left[\sum_{i=1}^{2} (x_i - y_i)^2\right]^{\frac{1}{2}} = \sqrt{(x_i - y_i)^2 + (x_2 - y_2)^2}$$

DE MANZANAS (CITY BLOCK):

$$d_2(\bar{x},\bar{y}) = \sum_{i=1}^{2} |x_i - y_i| = |x_i - y_i| + |x_2 - y_2|$$

CHEBYSHEV:

$$d_\infty(\bar{x},\bar{y}) = \max\{(\bar{x}_i - \bar{y}_i)\} = \max\{|x_i - y_i|, |x_2 - y_2|\}$$

CANBERRA:

$$d_c(\bar{x},\bar{y}) = \sum_{k=1}^{2} \frac{|x_{ik} - y_{jk}|}{(x_{ik} + y_{jk})} = \frac{|x_{i1} - y_{j1}|}{(x_{i1} + y_{j1})} + \frac{|x_{i2} - y_{j2}|}{(x_{i2} + y_{j2})}$$

$d_1 \rightarrow$ EUCLIDEA $\qquad \left[\sum_{i=1}^{2}\left(x_i - y_i\right)^2\right]^{1/2} = \sqrt{\left(x_1 - y_1\right)^2 + \left(x_2 - y_2\right)^2}$

	C 12-6	N 14-7	O 16-8	F 19-9	P 31-15	S 32-16	Cl. 35-17	Se. 79-34	Br. 80-35	Te. 128-52	I 127-53	At. 210-85
C 12 6	0	2,23	4,47	7,61	21,02	0,104	0,117	0,340	0,470	0,585	0,582	213,17
		0,010	0,020	0,035	0,098	22,36	25,07	72,61	100,28	124,78	124,23	1
N 14 7		0	2,23	5,38	18,78	20,12	23,25	69,64	71,69	122,56	122	210,95
			0,010	0,025	0,088	0,094	0,109	0,326	0,336	0,574	0,574	0,929
O 16 8			0	3,16	16,55	17,80	21,02	68,15	69,46	0,564	0,561	208,72
				0,014	0,077	0,083	0,098	0,319	0,325	120,33	119,77	0,979
F 19 9				0	13,41	14,76	17,88	65	66,30	117,17	116,61	205,56
					0,062	0,069	0,083	0,304	0,311	0,549	0,547	0,964
P 31 15					0	1,41	4,47	51,62	52,92	103,81	103,24	192,20
						0,006	0,020	0,242	0,248	0,486	0,484	0,901
S 32 16						0	3,16	50,32	51,62	102,52	101,95	190,9
							0,014	0,236	0,242	0,480	0,478	0,895
Cl. 35 17							0	47,16	48,46	99,36	98,43	187,74
								0,221	0,227	0,466	0,461	0,880
Se. 79 34								0	1,41	52,20	51,62	140,57
									0,006	0,244	0,242	0,659
Br. 80 35									0	0,238	0,236	139,28
										50,92	50,32	0,053
Te. 128 52										0	1,41	88,39
											0,006	0,414
I 127 53											0	88,95
												0,417
At. 210 85												0

$d_2 \longrightarrow$ CITY-BLOCK

$$\sum_{i=1}^{2} |x_i - y_i| = |x_1 - y_1| + |x_2 - y_2|$$

	C 12-6	N 14-7	O 16-8	F 19-9	P 31-15	S 32-16	Cl. 35-17	Se. 79-34	Br. 80-35	Te. 128-52	I 127-53	At. 210-85
C $^{12}_{6}$	0	3 0,010	6 0,021	10 0,036	28 0,101	30 0,108	34 0,122	95 0,342	97 0,350	162 0,584	162 0,58	277 1
N $^{14}_{7}$		0	3 0,010	7 0,625	25 0,090	27 0,097	31 0,111	92 0,332	94 0,339	159 0,574	159 0,574	274 0,989
O $^{16}_{8}$			0	4 0,014	22 0,079	24 0,086	28 0,101	89 0,321	91 0,328	156 0,563	156 0,563	271 0,978
F $^{19}_{9}$				0	18 0,064	20 0,072	24 0,086	85 0,306	87 0,314	152 0,548	152 0,548	267 0,963
P $^{31}_{15}$					0	2 0,007	6 0,021	67 0,241	69 0,249	134 0,483	134 0,483	249 0,898
S $^{32}_{16}$						0	4 0,014	65 0,234	67 0,241	132 0,476	132 0,476	247 0,891
Cl. $^{35}_{17}$							0	61 0,220	63 0,227	128 0,462	128 0,462	243 0,877
Se. $^{79}_{34}$								0	44 0,158	67 0,241	67 0,241	182 0,651
Br. $^{80}_{35}$									0	65 0,234	65 0,234	180 0,649
Te. $^{128}_{52}$										0	2 0,007	115 0,415
I $^{127}_{53}$											0	115 0,415
At. $^{210}_{85}$												0

$d_3 \longrightarrow L\infty \left(r=\infty \text{ en MINKOWSKI}\right) \text{ máx } \left(x_i - y_i\right) = \text{máx } \left\{\left(x_1 - y_1\right), \left|x_2 - y_2\right|\right\}$

	C 12-6	N 14-7	O 16-8	F 19-9	P 31-15	S 32-16	Cl. 35-17	Se. 79-34	Br. 80-35	Te. 128-52	I 127-53	At. 210-85
C 12 6	0	4 0,020	4 0,020	7 0,035	19 0,095	20 0,101	23 0,116	67 0,338	68 0,343	116 0,585	119 0,601	198 1
N 14 7		0	2 0,010	5 0,025	17 0,085	18 0,090	0,106 21	0,328 65	0,333 66	0,575 114	0,570 113	0,989 196
O 16 8			0	3 0,015	15 0,075	16 0,080	19 0,095	63 0,318	64 0,323	112 0,565	111 0,560	194 0,979
F 19 9				0	12 0,060	13 0,065	0,080 16	80 0,404	61 0,308	119 0,601	108 0,541	191 0,964
P 31 15					0	1 0,005	4 0,020	48 0,242	49 0,247	97 0,489	96 0,484	179 0,904
S 32 16						0	3 0,015	47 0,237	48 0,242	96 0,484	95 0,479	178 0,898
Cl. 35 17							0	44 0,222	45 0,227	93 0,469	92 0,464	175 0,883
Se. 79 34								0	1 0,005	49 0,247	48 0,242	131 0,661
Br. 80 35									0	48 0,242	47 0,237	130 0,656
Te. 128 52										0	1 0,005	82 0,414
I 127 53											0	83 0,419
At. 210 85												0

$d_4 \rightarrow$ CANBERRA $\quad dij = \sum_{k=1}^{2} \frac{|X_{ik} - Y_{jk}|}{(X_{ik} + Y_{jk})} = \frac{|X_{i1} - Y_{j1}|}{(X_{i1} + Y_{j1})} + \frac{|X_{i2} - Y_{j2}|}{(X_{i2} + Y_{j2})}$

	C 12-6	N 14-7	O 16-8	F 19-9	P 31-15	S 32-16	Cl. 35-17	Se. 79-34	Br. 80-35	Te. 128-52	I 127-53	At. 210-85
C 12 6	0	0,086 0,153	0,159 0,28	0,238 0,42	0,494 0,87	0,511 0,90	0,545 0,96	0,812 1,43	0,818 1,44	0,920 1,62	0,920 1,62	1 1,76
N 14 7		0	0,073 0,13	0,153 0,27	0,420 0,74	0,443 0,78	0,477 0,84	0,767 1,35	0,772 1,36	0,886 1,56	0,886 1,56	0,977 1,72
O 16 8			0	0,079 0,14	0,352 0,62	0,375 0,66	0,414 0,73	0,727 1,28	0,732 1,29	0,857 1,51	0,857 1,51	0,954 1,68
F 19 9				0	0,278 0,49	0,301 0,53	0,340 0,60	0,676 1,19	0,681 1,20	0,818 1,44	0,818 1,44	0,931 1,64
P 31 15					0	0,272 0,48	0,068 0,12	0,465 0,82	0,477 0,84	0,659 1,16	0,659 1,16	0,818 1,44
S 32 16						0	0,039 0,07	0,443 0,78	0,454 0,8	0,636 1,12	0,642 1,13	0,801 1,41
Cl. 35 17							0	0,403 0,71	0,414 0,73	0,607 1,07	0,613 1,08	0,784 1,38
Se. 79 34								0	0,011 0,02	0,250 0,44	0,255 0,45	0,465 0,82
Br. 80 35									0	0,238 0,42	0,244 0,43	0,488 0,86
Te. 128 52										0	0,005 0,01	0,437 0,77
I 127 53											0	0,267 0,47
At. 210 85												0

2. En segundo lugar, hemos normalizado todos los elementos de las matrices anteriores siguiendo el siguiente proceso:

A)

$$S_o(d_1) = \delta_o(d_1, \bar{0}) = \max \left\{ \left| d_1(\bar{x},\bar{y}) - \bar{0}(\bar{x},\bar{y}) \right| : \bar{x},\bar{y} \in X \right\}$$

$$S_o(d_2) = \delta_o(d_2, \bar{0}) = \max \left\{ \left| d_2(x,y) - 0(x,y) \right| : x,y \in X \right\}$$

$$S_o(d_\alpha) = \delta_o(d_\alpha, 0) = \max \left\{ \left| d_\alpha(x,y) - 0(x,y) \right| : x,y \in X \right\}$$

$$S_o(d_c) = \delta_o(d_c, 0) = \max \left\{ \left| d_c(x,y) - 0(x,y) \right| : x,y \in X \right\}$$

B)
$$\delta'_o(d_i,d_j) = \delta_o\left(d_i/_{S_o(d_i)}, d_j/_{S_o(d_j)}\right) =$$

$$= \max \left| \frac{d_i(x,y)}{S_o(d_i)} - \frac{d_j(x,y)}{S_o(d_j)} \right| : x,y \in X$$

y obtenemos las siguientes matrices:

000269

$d_1 \rightarrow$ EUCLIDEA

$$\left[\sum_{i=1}^{2}\left(x_i - y_i\right)^2\right]^{1/2} = \sqrt{\left(x_1 - y_1\right)^2 + \left(x_2 - y_2\right)^2}$$

	C 12-6	N 14-7	O 16-8	F 19-9	P 31-15	S 32-16	Cl. 35-17	Se. 79-34	Br. 80-35	Te. 128-52	I 127-53	At. 210-85
C 12 6	0	2,23	4,47	7,61	21,02	0,104	0,117	0,340	0,470	0,585	0,582	213,17
		0,010	0,020	0,035	0,098	22,36	25,07	72,61	100,28	124,78	124,23	1
N 14 7		0	2,23	5,38	18,78	20,12	23,25	69,64	71,69	122,56	122	210,95
			0,010	0,025	0,088	0,094	0,109	0,326	0,336	0,574	0,574	0,929
O 16 8			0	3,16	16,55	17,80	21,02	68,15	69,46	0,564	0,561	208,7
				0,014	0,077	0,083	0,098	0,319	0,325	120,33	119,77	0,979
F 19 9				0	13,41	14,76	17,88	65	66,30	117,17	116,61	20
					0,062	0,069	0,083	0,304	0,311	0,549	0,547	0,964
P 31 15					0	1,41	4,47	51,62	52,92	103,81	103,24	192,20
						0,006	0,020	0,242	0,248	0,486	0,484	0,901
S 32 16						0	3,16	50,32	51,62	102,52	101,95	190,9
							0,014	0,236	0,242	0,480	0,478	0,895
Cl. 35 17							0	47,16	48,46	99,36	98,43	187,74
								0,221	0,227	0,466	0,461	0,880
Se. 79 34								0	1,41	52,20	51,62	140,57
									0,006	0,244	0,242	0,659
Br. 80 35									0	0,238	0,236	139,28
										50,92	50,32	0,053
Te. 128 52										0	1,41	88,39
											0,006	0,414
I 127 53											0	88,95
												0,417
At. 210 85												0

$d_2 \longrightarrow$ CITY-BLOCK

$$\sum_{i=1}^{2} |x_i - y_i| = |x_1 - y_1| + |x_2 - y_2|$$

	C 12-6	N 14-7	O 16-8	F 19-9	P 31-15	S 32-16	Cl. 35-17	Se. 79-34	Br. 80-35	Te. 128-52	I 127-53	At. 210-85
C $^{12}_{6}$	0	3 0,010	6 0,021	10 0,036	28 0,101	30 0,108	34 0,122	95 0,342	97 0,350	162 0,584	162 0,58	277 1
N $^{14}_{7}$		0	3 0,010	7 0,625	25 0,090	27 0,097	31 0,111	92 0,332	94 0,339	159 0,574	159 0,574	274 0,989
O $^{16}_{8}$			0	4 0,014	22 0,079	24 0,086	28 0,101	89 0,321	91 0,328	156 0,563	156 0,563	271 0,978
F $^{19}_{9}$				0	18 0,064	20 0,072	24 0,086	85 0,306	87 0,314	152 0,548	152 0,548	267 0,963
P $^{31}_{15}$					0	2 0,007	6 0,021	67 0,241	69 0,249	134 0,483	134 0,483	249 0,898
S $^{32}_{16}$						0	4 0,014	65 0,234	67 0,241	132 0,476	132 0,476	247 0,891
Cl. $^{35}_{17}$							0	61 0,220	63 0,227	128 0,462	128 0,462	243 0,877
Se. $^{79}_{34}$								0	44 0,158	67 0,241	67 0,241	182 0,651
Br. $^{80}_{35}$									0	65 0,234	65 0,234	180 0,649
Te. $^{128}_{52}$										0	2 0,007	115 0,415
I $^{127}_{53}$											0	115 0,415
At. $^{210}_{85}$												0

$d_3 \longrightarrow L\infty \left(r=\infty \text{ en MINKOWSKI}\right) \text{ máx } \left(x_i - y_i\right) = \text{máx}\left\{\left(x_1 - y_1\right), \left|x_2 - y_2\right|\right\}$

	C 12-6	N 14-7	O 16-8	F 19-9	P 31-15	S 32-16	Cl. 35-17	Se. 79-34	Br. 80-35	Te. 128-52	I 127-53	At. 210-85
C 12 6	0	4 0,020	4 0,020	7 0,035	19 0,095	20 0,101	23 0,116	67 0,338	68 0,343	116 0,585	119 0,601	198 1
N 14 7		0	2 0,010	5 0,025	17 0,085	18 0,090	0,106 21	0,328 65	0,333 66	0,575 114	0,570 113	0,989 196
O 16 8			0	3 0,015	15 0,075	16 0,080	19 0,095	63 0,318	64 0,323	112 0,565	111 0,560	194 0,979
F 19 9				0	12 0,060	13 0,065	0,080 16	80 0,404	61 0,308	119 0,601	108 0,541	191 0,964
P 31 15					0	1 0,005	4 0,020	48 0,242	49 0,247	97 0,489	96 0,484	179 0,904
S 32 16						0	3 0,015	47 0,237	48 0,242	96 0,484	95 0,479	178 0,898
Cl. 35 17							0	44 0,222	45 0,227	93 0,469	92 0,464	175 0,883
Se. 79 34								0	1 0,005	49 0,247	48 0,242	131 0,661
Br. 80 35									0	48 0,242	47 0,237	130 0,656
Te. 128 52										0	1 0,005	82 0,414
I 127 53											0	83 0,419
At. 210 85												0

$d_4 \rightarrow$ CANBERRA $\quad d_{ij} = \sum_{k=1}^{2} \frac{|X_{ik} - Y_{jk}|}{(X_{ik} + Y_{jk})} = \frac{|X_{i1} - Y_{j1}|}{(X_{i1} + Y_{j1})} + \frac{|X_{i2} - Y_{j2}|}{(X_{i2} + Y_{j2})}$

	C 12-6	N 14-7	O 16-8	F 19-9	P 31-15	S 32-16	Cl. 35-17	Se. 79-34	Br. 80-35	Te. 128-52	I 127-53	At. 210-85
C 12 6	0	0,086 0,153	0,159 0,28	0,238 0,42	0,494 0,87	0,511 0,90	0,545 0,96	0,812 1,43	0,818 1,44	0,920 1,62	0,920 1,62	1 1,76
N 14 7		0	0,073 0,13	0,153 0,27	0,420 0,74	0,443 0,78	0,477 0,84	0,767 1,35	0,772 1,36	0,886 1,56	0,886 1,56	0,977 1,72
O 16 8			0	0,079 0,14	0,352 0,62	0,375 0,66	0,414 0,73	0,727 1,28	0,732 1,29	0,857 1,51	0,857 1,51	0,954 1,68
F 19 9				0	0,278 0,49	0,301 0,53	0,340 0,60	0,676 1,19	0,681 1,20	0,818 1,44	0,818 1,44	0,931 1,64
P 31 15					0	0,272 0,48	0,068 0,12	0,465 0,82	0,477 0,84	0,659 1,16	0,659 1,16	0,818 1,44
S 32 16						0	0,039 0,07	0,443 0,78	0,454 0,8	0,636 1,12	0,642 1,13	0,801 1,41
Cl. 35 17							0	0,403 0,71	0,414 0,73	0,607 1,07	0,613 1,08	0,784 1,38
Se. 79 34								0	0,011 0,02	0,250 0,44	0,255 0,45	0,465 0,82
Br. 80 35									0	0,238 0,42	0,244 0,43	0,488 0,86
Te. 128 52										0	0,005 0,01	0,437 0,77
I 127 53											0	0,267 0,47
At. 210 85												0

3.- EN TERCER LUGAR OBTENEMOS LA MATRIZ CORRESPONDIENTE A LA METRICA δ^* DE LAS METRICAS YA INDICADAS :

$$\delta^*(d_i, d_j) = \delta^*\left(d_i/\int_o(d_i),\ d_j/\int_o(d_j)\right) =$$

$$= \max\left\{\left|\frac{d_i(x.y)}{\int_o(d_i)} - \frac{d_j(x.y)}{\int_o(d_j)}\right| : x, y \in P\right\}$$

DONDE i Y j VARIAN ENTRE 1,2,3 y 4

δ^*	d_1	d_2	d_3	d_4
d_1	0	0,152	0,127	0,472
d_2		0	0,153	0,470
d_3			0	0,475
d_4				0

UNA VEZ OBTENIDA LA MATRIZ ANTERIOR UTILIZAMOS A PARTIR DE ELLA, COMO OTUS A d_1, d_2, d_3 y d_4 UTILIZAMOS TRES METODOS JERARQUICOS DEL ANALISIS DE CONGLOMERADOS Y CONSTRUIMOS LOS CORRESPONDIENTES DENDROGRAMAS; LOS METODOS UTILIZADOS SON ENLACE SENCILLO, DISTANCIA PROMEDIO NO PONDERADO Y ENLACE COMPLETO.

1.- ENLACE SENCILLO (DISTANCIA MINIMA)

δ^*	d_1	d_2	d_3	d_4
d_1	0	0,152	0,127	0,472
d_2		0	0,153	0,470
d_3			0	0,475
d_4				0

d_1 y d_2 SON FUSIONADOS PARA FORMAR EL PRIMER CONGLOMERADO PUESTO QUE δ^*_{13} ES EL MAS PEQUEÑO, 0,127, DE LOS ELEMENTOS DE LA MATRIZ (ES LA MENOR DISTANCIA, CON LA NUEVA METRICA, ENTRE ELEMENTOS DEL CONJUNTO $\{d_1, d_2, d_3, d_4\}$)
LAS DISTANCIAS EN ESTE NUEVO GRUPO (d_1, d_3) Y LOS RESTANTES UTOS, d_2 y d_4 SE OBTIENEN A PARTIR DE LA MATRIZ DE DISTANCIAS DE LA SIGUIENTE MANERA:

$$\delta^*(d_1, d_3)d_2 = \min\{d^*_{12}, d^*_{32}\} = \min\{0{,}152, 0{,}153\} = 0{,}152$$

$$\delta^*(d_1, d_3)d_4 = \min\{d^*_{14}, d^*_{34}\} = \min\{0{,}472, 0{,}475\} = 0{,}472$$

Y FORMAMOS UNA NUEVA MATRIZ DE DISTANCIAS δ_1^* CON LAS DISTANCIAS ENTRE INDIVIDUOS Y ENTRE GRUPO - INDIVIDUOS :

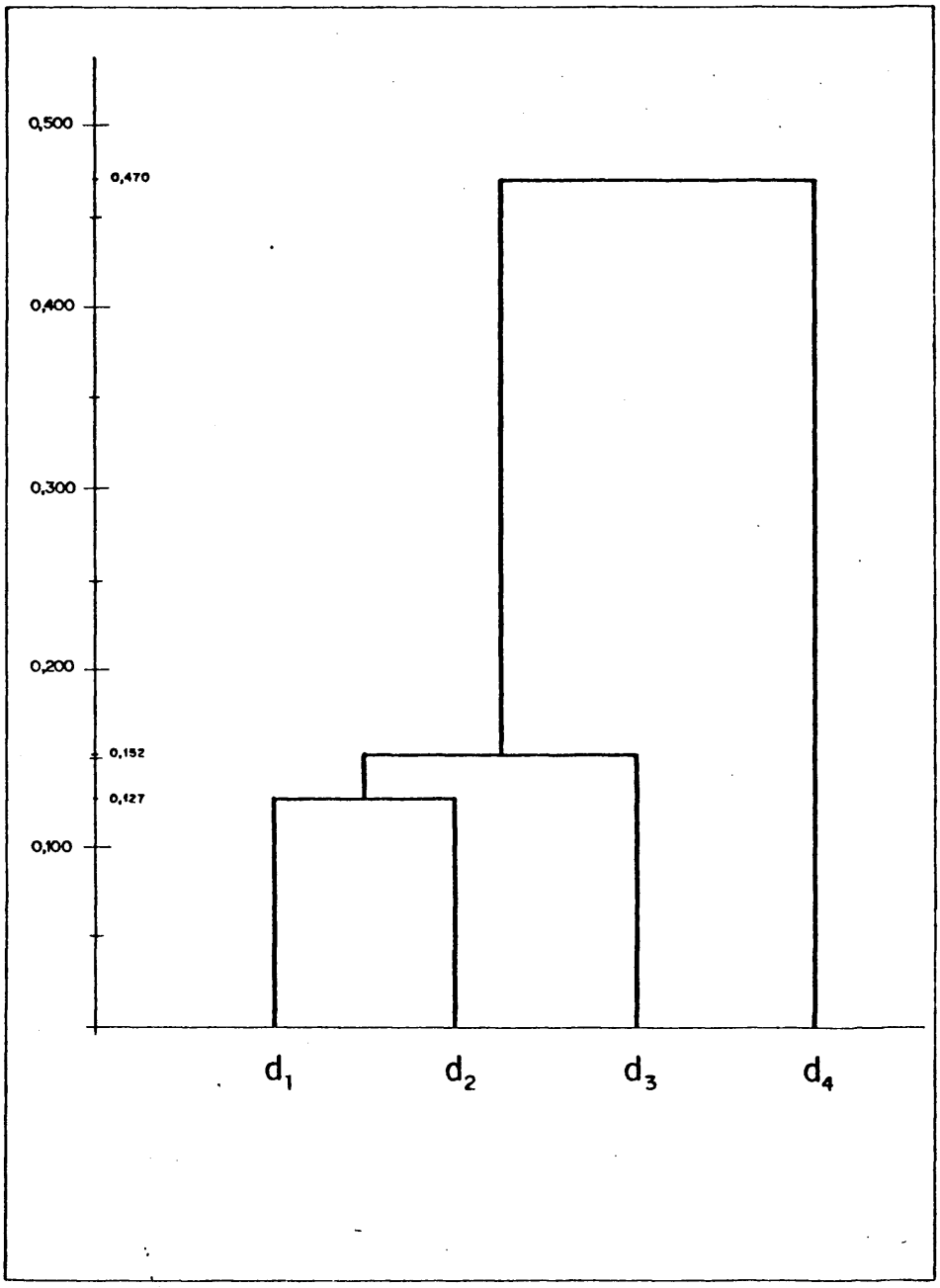
	$(d_1\ d_3)$	d_2	d_4
$(d_1\ d_3)$	0	0,152	0,472
d_2		0	0,470
d_4			0

COMO EL ELEMENTO MAS PEQUEÑO DE LA MATRIZ ES 0,152 , SE FORMA UN NUEVO GRUPO $(d_1\ d_3, d_2)$ EL OTU d_2 SE AÑADE AL GRUPO QUE FORMAN LOS OTUS d_1 Y d_3

LA DISTANCIA, δ^*, ENTRE EL NUEVO GRUPO Y EL OTU d_4 ES EL SIGUIENTE :

$$\delta^*\left[(d_1\ d_3), d_2\right], d_4\} = \min\left\{\delta^*\left[(d_1\ d_3)d_4\right]\ \delta^*(d_2, d_4)\right\} =$$

= min 0,472 , 0,470 = 0,470 , YA ESTAN AGRUPADOS O CONGLOMERADOS LOS CUATRO OTUS , Y EL CORRESPONDIENTE DENDROGRAMA ES :


0,500
0,470
0,400
0,300
0,200
0,152
0,127
0,100
d1
d2
d3
d4

2.- DISTANCIA PROMEDIO NO PONDERADO

δ^*	d_1	d_2	d_3	d_4
d_1	0	0,152	0,127	0,477
d_2		0	0,153	0,470
d_3			0	0,475
d_4				0

SE CONSIDERA QUE LA DISTANCIA ENTRE DOS GRUPOS VIENE DEFINIDA POR EL PROMEDIO DE LAS DISTANCIAS DE LOS ELEMENTOS DE UN GRUPO CON RESPECTO A LOS DEL OTRO.
PARTIMOS DEL NIVEL 0 Y HEMOS DE OBTENER EL NIVEL 1 A BASE DE AGRUPAR LOS DOS ELEMENTOS CUYA DISTANCIA SEA MINIMA, QUE EN ESTE CASO SON d_1 Y d_3

$$\delta^* (d_1, d_3) = 0,127$$

LUEGO EL PRIMER CONGLOMERADO ES (d_1, d_3)
OBTENEMOS A CONTINUACION LAS DISTANCIAS, δ^*, ENTRE EL GRUPO ANTERIOR Y LOS OTUS d_2 Y d_4 :

$$\delta^* \left[(d_1, d_3), d_2 \right] = \frac{\delta^*(d_1, d_2) + \delta^*(d_3, d_2)}{2} = 0,1525$$

$$\delta^* \left[(d_1, d_3), d_4 \right] = \frac{\delta^*(d_1, d_4) + \delta^*(d_3, d_4)}{2} = 0,4735$$

LO QUE NOS PROPORCIONA LA SIGUIENTE MATRIZ :

	$(d_1\ d_3)$	d_2	d_4
$(d_1\ d_3)$	0	0,1525	0,4735
d_2		0	0,470
d_4			0

COMO LOS ELEMENTOS QUE MENOR DISTANCIA, δ^*, TIENEN ENTRE SI SON (d_1, d_3) Y d_2 EL ELEMENTO d SE FUSIONA CON EL CONGLOMERADO (d_1, d_3) (AL NIVEL DE DISTANCIA 0,1525)

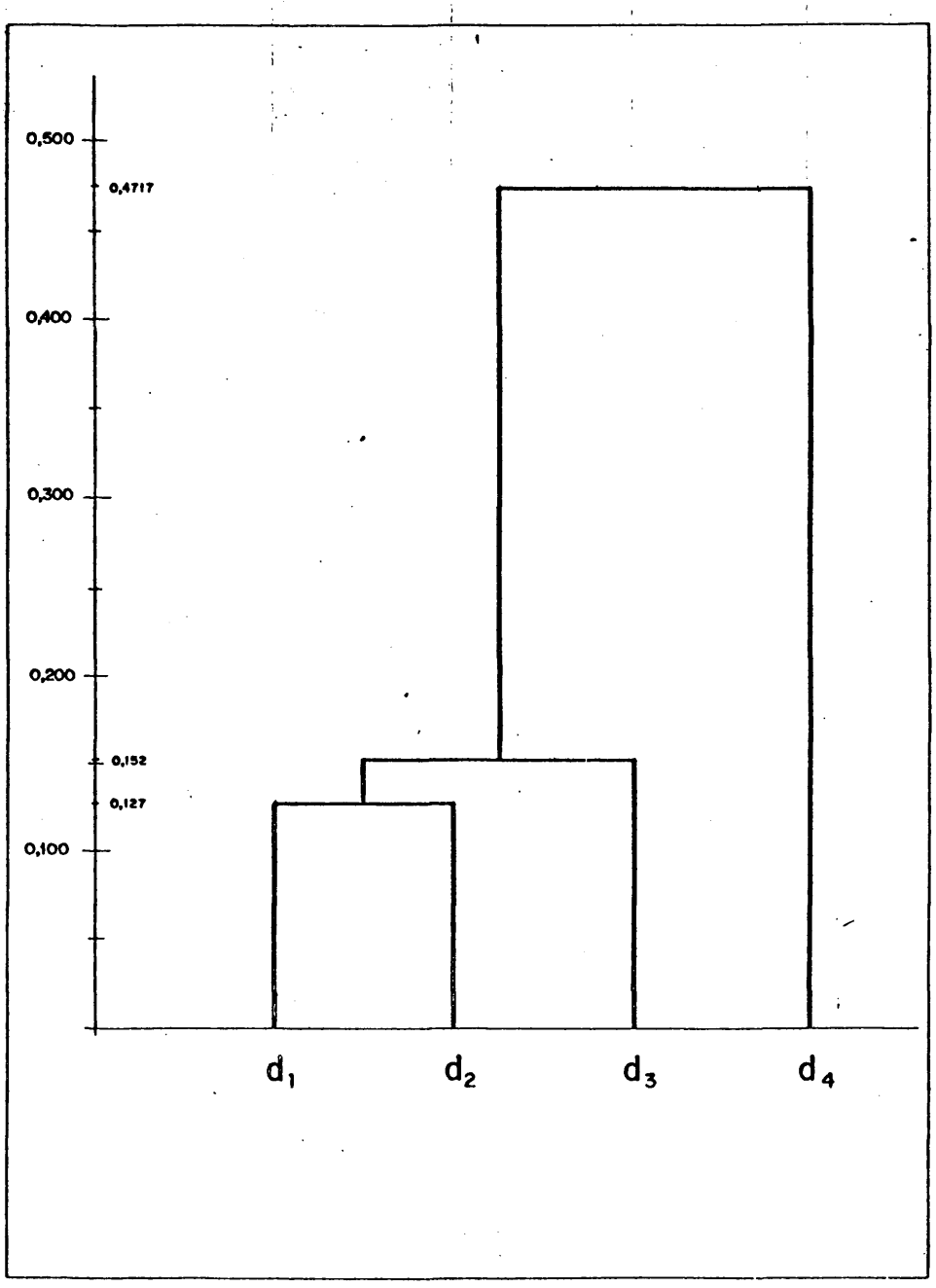
Y, FINALMENTE, HEMOS DE HALLAR LA DISTANCIA, δ^*, ENTRE EL NUEVO CONGLOMERADO

$$\left[(d_1, d_3), d_2\right] \quad \text{Y EL OTU } d_4 :$$

$$\delta^*\left\{\left[(d_1, d_3), d_2\right] d_4\right\} = \frac{\delta^*\left[(d_1, d_3), d_4\right] + \delta^*\left[d_2, d_4\right]}{2} = 0{,}4717$$

POR TANTO EL CORRESPONDIENTE DENDROGRAMA ES:

000279



3.- ENLACE COMPLETO (DISTANCIA MAXIMA)

δ^*	d_1	d_2	d_3	d_4
d_1	0	0,152	0,127	0,472
d_2		0	0,153	0,470
d_3			0	0,475
d_4				0

COMENZAMOS, AL IGUAL QUE EN EL MÉTODO DE LA UNIÓN SENCILLA, FUSIONANDO LOS INDIVIDUOS d Y d Y OBTENEMOS EL CONGLOMERADO (d , d)
LA DISTANCIA ENTRE ESTE CONGLOMERADO Y LOS DOS RESTANTES OTUS SE OBTIENE A PARTIR DE LA MATRIZ DE DISTANCIA, δ^*, DE LA SIGUIENTE FORMA :

$$\delta^*\left[(d_1, d_3)\, d_2\right] = \text{máx}\left\{\delta^*(d_1, d_2), \delta^*(d_3, d_2)\right\} = \text{máx}\left\{0{,}152, 0{,}153\right\} = 0{,}153$$

$$\delta^*\left[(d_1, d_3)\, d_4\right] = \text{máx}\left\{\delta^*(d_1, d_4), \delta^*(d_3, d_4)\right\} = \text{máx}\left\{0{,}472, 0{,}475\right\} = 0{,}475$$

FORMAMOS LA NUEVA MATRIZ DE DISTANCIAS :

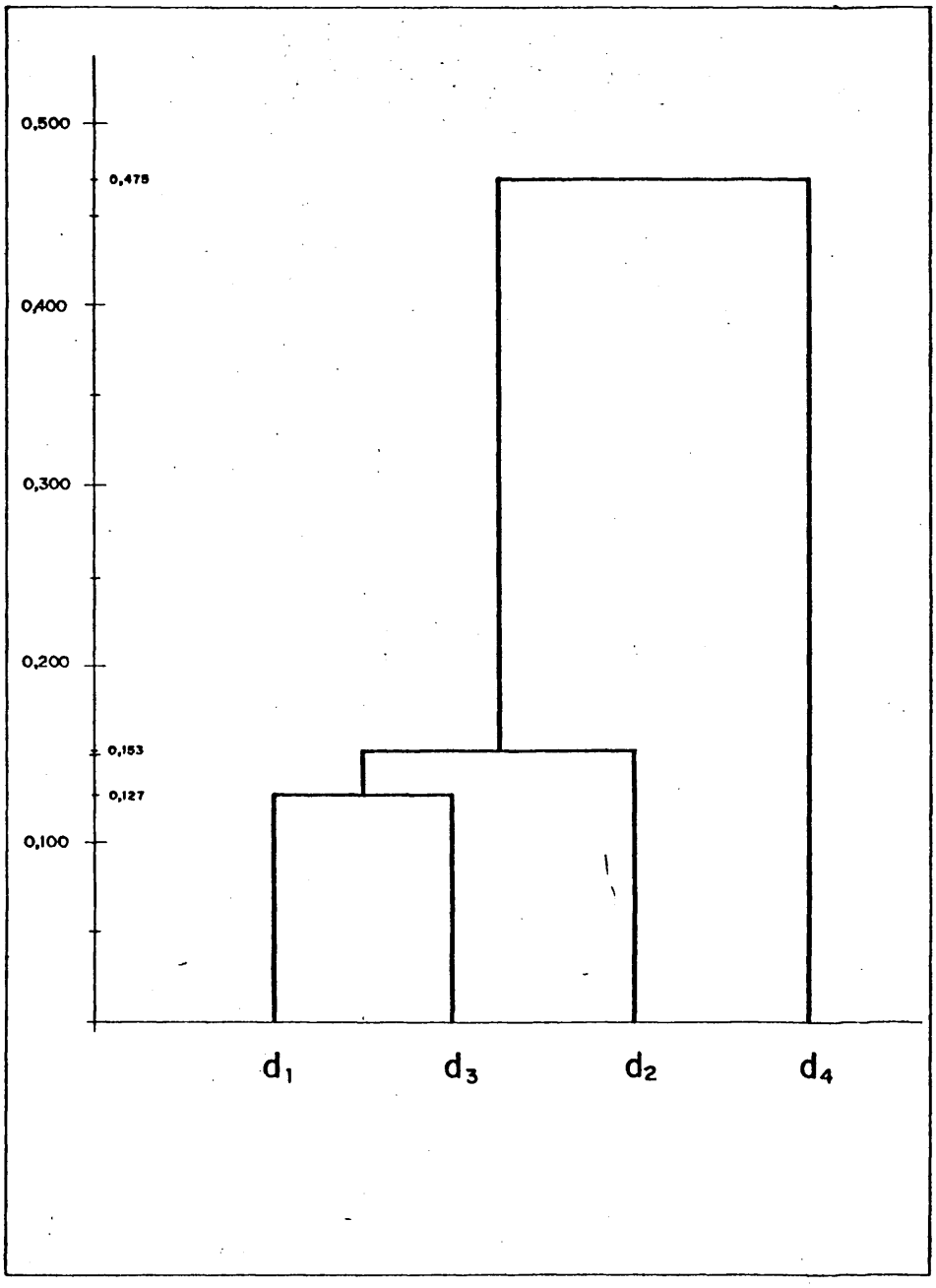
δ^*	$(d_1\ d_3)$	d_2	d_4
$(d_1\ d_3)$	0	0,153	0,475
d_2		0	0,470
d_4			0

COMO EL NUMERO MAS PEQUEÑO DE ESTA MATRIZ ES 0,153 SE FORMA UN NUEVO GRUPO $\left[(d_1\ d_3),\ d_2\right]$, LA DISTANCIA, δ^*, ENTRE EL NUEVO GRUPO $\left[(d_1\ d_3), d_2\right]$ Y EL OTU d_4 ES:

$$\delta^*\left\{\left[(d_1\ d_3), d_2\right], d_4\right\} = \text{máx}\left\{\delta^*\left[(d_1, d_3), d_4\right],\right.$$

$$\left.\delta^*\left[d_2, d_4\right]\right\} = \text{máx}\left\{0{,}475\ ,\ 0{,}470\right\} = 0{,}475$$

Y EL DENDROGRAMA CORRESPONDIENTE ES :


0,500
0,475
0,400
0,300
0,200
0,153
0,127
0,100
d1
d3
d2
d4

D. RESULTADOS

A la vista de los dendrogramas anteriores, ya podemos deducir consecuencias:

a. Las métricas más afines son las d_1 y d_3. Euclidea y Chebyshev.

b. Las métricas más dispares son d_3 y d_4, Chebyshev y Canberra.

c. La métrica de Canberra es la métrica que más discrepa de las demás.

d. La métrica de Canberra es la más costosa en tiempo de cálculo y la más difícil de llevar a cabo.

e. La métrica d_2, "de manzanas" (CITY BLOCK), es intermedia entre d_1 y d_3 por un lado y d_4 por otro.

COEFICIENTE DE CORRELACION COFENETICO

Para medir el grado de concordancia entre los dendrogramas obtenidos por diferentes métodos de conglomeración, utilizamos el método de las CORRELACIONES COFENETICAS de SOKAL y ROHLF.

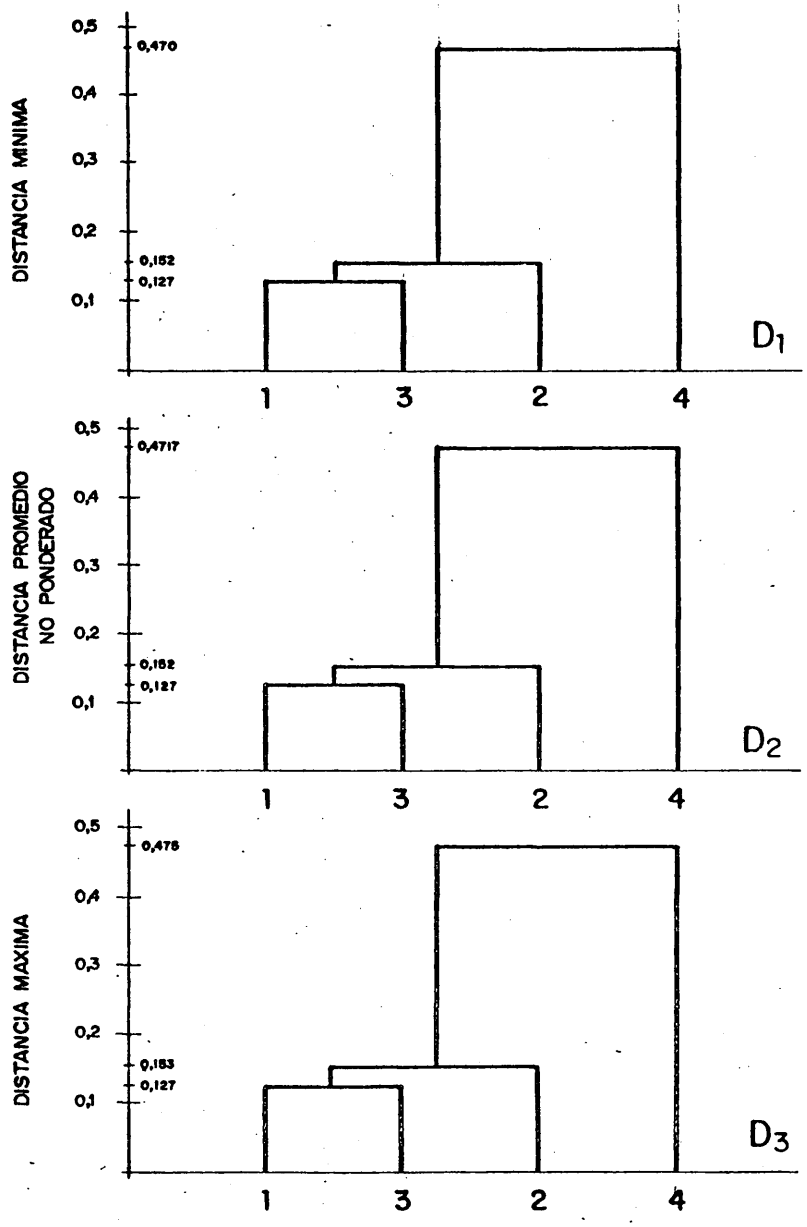
Se define como valor cofenético de dos elementos, el número que expresa el nivel al que dichos elementos se unen en un dendrograma.

A partir de un dendrograma se obtiene la correspondiente matriz de valores cofenéticos.

La comparación de dos dendrogramas mediante este método, se hace calculando el coeficiente de correlación producto-momento entre las correspondientes matrices de valores cofenéticos. Digamos que los autores de este método lo utilizaron para medir la correlación entre los datos originales y los resultados de un dendrograma.

La magnitud del coeficiente de correlación así obtenido, que se denomina coeficiente de correlación cofenético describirá la concordancia entre los dendrogramas comparados y, por tanto, entre los métodos de conglomeración correspondientes.

DENDROGRAMAS A COMPARAR:



MATRICES DE VALORES COFENÉTICOS

$D_1 \longrightarrow$

	d_1	d_2	d_3	d_4
d_1	X	0,152	0,127	0,47
d_2		X	0,152	0,47
d_3			X	0,47
d_4				X

$D_2 \longrightarrow$

	d_1	d_2	d_3	d_4
d_1	X	0,152	0,127	0,4717
d_2		X	0,152	0,4717
d_3			X	0,4717
d_4				X

$D_3 \longrightarrow$

	d_1	d_2	d_3	d_4
d_1	X	0,153	0,127	0,475
d_2		X	0,153	0,475
d_3			X	0,475
d_4				X

COMPARACION DE LOS DENDROGRAMAS D_1 y D_2:

Se hace utilizando las matrices de valores cofenéticos y calculando un coeficiente de correlación entre los dos conjuntos de elementos de dichas matrices.

Utilizaremos el coeficiente de correlación de PEARSON:

$$r_{jk} = \frac{\sum_{i=1}^{n} X_{ij} X_{ik} - \frac{1}{n} \left(\sum_{i=1}^{n} X_{ij}\right) \left(\sum_{i=1}^{n} X_{ik}\right)}{\left\{\left[\sum_{i=1}^{n} X_{ij}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} X_{ij}\right)^2\right]\left[\sum_{i=1}^{n} X_{ik}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} X_{ik}\right)\right]\right\}^{\frac{1}{2}}}$$

o bien la fórmula:

$$r_{jk} = \frac{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - X_k)}{\left\{\sum_{i=1}^{n} (X_{ij} - X_j)^2 \sum_{i=1}^{n} (X_{ik} - X_k)^2\right\}^{\frac{1}{2}}}$$

Los coeficientes de correlación cofenéticos que hemos obtenido al comparar las matrices de valores cofenéticos asociados, respectivamente, a los dendrogramas anteriores son,

$$r_{D_1 D_2} = 0{,}840$$

$$r_{D_1 D_3} = 0{,}912$$

$$r_{D_2 D_3} = 0{,}972$$

Por tanto, la matriz de coeficientes de correlación cofenéticos es

r	D_1	D_2	D_3
D_1		0,840	0,912
D_2			0,972
D_3			

Si observamos los coeficientes de correlación cofenéticos que hemos obtenido, podemos hablar de la mayor o menor similitud (o disimilitud) entre los dendrogramas obtenidos y, por tanto, de los métodos de conglomeración correspondientes.

Los métodos más similares son los de la distancia promedio no ponderado, y el de la distancia máxima, y los menos similares son los métodos de la distancia mínima y de la distancia promedio no ponderado.

RESOLUCION DEL PROBLEMA ANTERIOR SOBRE EJEMPLARES DE NEUROPTERIS

A continuación, exponemos de una forma similar y totalmente paralela a la realizada anteriormente, la resolución del problema del estudio comparativo de métricas, pero tomando como conjunto inicial de OTUS los trece ejemplares de NEUROPTERIS, que hemos clasificado en el capítulo anterior.

La exposición será esquemática y en el mismo orden. Ofrecemos, únicamente, los resultados obtenidos.

d_1 =Euclidea	A	B	C	D	E	F	G	H	I	J	K	L	M
	3	3,2	2,3	2,6	3,7	3,9	4,9	6,6	6,9	5,7	4,8	4,6	7,8
	2,1	2,2	1,3	1,2	1,3	1,7	1,4	3,3	3,7	2,4	1,6	2,2	2,6
	1,2	1,4	1,4	1,2	1,7	1,2	0,7	2	2,1	1,1	1,4	1,8	1,4
	2,3	2,3	1,8	2	2,8	2,7	4,1	5,8	6,1	5	3,9	3,8	6,6
A 3-2,1-1,2-2,3	0	0,3 0,03	1,19 0,15	1,04 0,13	1,27 0,16	1,06 0,13	2,69 0,34	5,22 0,66	5,74 0,73	3,82 0,48	2,46 0,31	2,07 0,26	6,46 0,82
B 3,2-2,2-1,4-2,3		0	1,36 0,17	1,22 0,15	1,1 0,14	0,96 0,12	2,69 0,34	5,03 0,64	6,2 0,79	3,69 0,47	2,34 0,29	2,09 0,26	6,30 0,80
C 2,3-1,3-1,4-1,8			0	0,42 0,05	1,74 0,22	1,88 0,23	3,54 0,45	5,98 0,76	6,77 0,86	4,76 0,60	3,27 0,41	3,20 0,40	7,41 0,94
D 2,6-1,2-1,2-2				0	1,45 0,18	1,55 0,19	3,16 0,40	5,95 0,75	6,91 0,88	4,47 0,57	2,94 0,37	2,93 0,37	7,84 1
E 3,7-1,3-1,7-2,8					0	0,67 0,08	2,03 0,25	4,63 0,59	5,24 0,66	3,22 0,41	1,61 0,20	1,62 0,20	5,61 0,71
F 3,9-1,7-1,2-2,7						0	1,81 0,23	4,48 0,57	5,03 0,64	3 0,38	1,51 0,19	1,51 0,19	5,59 0,71
G 4,9-1,4-0,7-4,1							0	3,07 0,391	3,9 0,49	1,61 0,20	1,78 0,22	1,42 0,18	4,07 0,51
H 6,6-3,3-2-5,8								0	0,53 0,06	1,75 0,22	3,17 0,40	3,04 0,38	1,71 0,21
I 6,9-3,7-2,1-6,1									0	2,31 0,29	3,7 0,47	3,59 0,45	1,66 0,21
J 5,7-2,4-1,1-5										0	1,65 0,21	1,78 0,22	2,66 0,33
K 4,8-1,6-1,4-3,9											0	0,75 0,04	4,15 0,52
L 4,6-2,2-1,8-3,8												0	4,28 0,54
M 7,8-2,6-1,4-6,6													0

MATRIZ DE DISTANCIAS EUCLIDEAS ENTRE 13 EJEMPLARES DE NEUROPTERIS

d_2= CITY-BLOCK O DE MANZANAS	A	B	C	D	E	F	G	H	I	J	K	L	M
	3	3,2	2,3	2,6	3,7	3,9	4,9	6,6	6,9	5,7	4,8	4,6	7,8
	2,1	2,2	1,3	1,2	1,3	1,7	1,4	3,3	3,7	2,4	1,6	2,2	2,6
	1,2	1,4	1,4	1,2	1,7	1,2	0,7	2	2,1	1,1	1,4	1,8	1,4
	2,3	2,3	1,8	2	2,8	2,7	4,1	5,8	6,1	5	3,9	3,8	6,6
A 3-2,1-1,2-2,3	0	0,5 0,03	2,2 0,16	1,6 0,11	2,5 0,18	1,7 0,12	4,9 0,36	8,5 0,62	10,2 0,75	5,8 0,42	4,1 0,30	3,8 0,27	13,6 1
B 3,2-2,2-1,4-2,3		0	2,3 0,16	2,1 0,15	2,2 0,16	1,8 0,13	5 0,36	8,6 0,63	9,7 0,71	5,7 0,41	3,8 0,27	3,3 0,24	9,3 0,68
C 2,3-1,3-1,4-1,8			0	0,8 0,05	2,7 0,19	3,1 0,22	5,7 0,41	10,9 0,80	12 0,88	8 0,58	4,9 0,36	5,6 0,41	11,4 0,83
D 2,6-1,2-1,2-2				0	2,5 0,18	2,5 0,18	5,1 0,37	10,7 0,78	11,8 0,86	7,4 0,54	4,7 0,34	5,4 0,34	11,4 0,83
E 3,7-1,3-1,7-2,8					0	1,2 0,08	3,6 0,26	8,2 0,60	9,3 0,68	5,9 0,43	2,8 0,20	2,9 0,21	9,5 0,69
F 3,9-1,7-1,2-2,7						0	3,2 0,23	8,2 0,60	9,3 0,68	4,9 0,36	2,4 0,17	2,9 0,21	8,9 0,65
G 4,9-1,4-0,7-4,1							0	6,6 0,48	6,7 0,49	3,1 0,22	1,2 0,08	2,5 0,18	7,3 0,53
H 6,6-3,3-2-5,8								0	1,1 0,08	3,5 0,25	6 0,44	5,3 0,38	3,3 0,24
I 6,9-3,7-2,1-6,1									0	4,6 0,33	6,9 0,50	6,4 0,47	3,2 0,23
J 5,7-2,4-1,1-5										0	3,1 0,22	3 0,22	4,2 0,30
K 4,8-1,6-1,4-3,9											0	5,5 0,404	6,7 0,49
L 4,6-2,2-1,8-3,8												0	6,8 0,5
M 7,8-2,6-1,4-6,6													0

MATRIZ DE DISTANCIAS DE MANZANAS (CITY-BLOCK) ENTRE 13 EJEMPLARES DE NEUROPTERIS

d_3 = CHEBSCHEV	A	B	C	D	E	F	G	H	I	J	K	L	M
	3	3,2	2,3	2,6	3,7	3,9	4,9	6,6	6,9	5,7	4,8	4,6	7,8
	2,1	2,2	1,3	1,2	1,3	1,7	1,4	3,3	3,7	2,4	1,6	2,2	2,6
	1,2	1,4	1,4	1,2	1,7	1,2	0,7	2	2,1	1,1	1,4	1,8	1,4
	2,3	2,3	1,8	2	2,8	2,7	4,1	5,8	6,1	5	3,9	3,8	6,6
A 3-2,1-1,2-2,3	0	0,2 0,03	0,8 0,15	0,9 0,16	0,8 0,15	0,9 0,16	1,9 0,35	3,6 0,67	3,9 0,73	2,7 0,50	1,8 0,33	1,6 0,30	4,8 0,90
B 3,2-2,2-1,4-2,3		0	0,9 0,16	1 0,18	0,9 0,16	0,7 0,13	1,8 0,33	3,5 0,66	3,8 0,71	2,7 0,50	1,6 0,30	1,5 0,28	4,6 0,86
C 2,3-1,3-1,4-1,8			0	0,3 0,05	1,4 0,26	1,6 0,30	2,6 0,49	4,3 0,81	4,6 0,86	3,4 0,64	2,5 0,47	2,3 0,43	4,8 0,90
D 2,6-1,2-1,2-2				0	1,8 0,33	1,3 0,24	1,3 0,24	4 0,75	4,3 0,81	4,1 0,77	2,2 0,41	2 0,37	5,3 1
E 3,7-1,3-1,7-2,8					0	0,5 0,09	1,3 0,24	3 0,56	3,3 0,62	2,2 0,41	1,1 0,20	1 0,18	4,1 0,77
F 3,9-1,7-1,2-2,7						0	1,4 0,26	3,1 0,58	3,4 0,64	2,3 0,43	1,2 0,22	1,2 0,22	3,9 0,73
G 4,9-1,4-0,7-4,1							0	1,9 0,35	2 0,37	1 0,18	0,7 0,13	1,1 0,20	2,9 0,54
H 6,6-3,3-2-5,8								0	0,4 0,07	0,9 0,16	1,9 0,35	2 0,37	1,2 0,22
I 6,9-3,7-2,1-6,1									0	1,3 0,24	2,2 0,41	2,3 0,43	0,9 0,16
J 5,7-2,4-1,1-5										0	1,1 0,20	1,2 0,22	2,1 0,39
K 4,8-1,6-1,4-3,9											0	0,6 0,113	3 0,56
L 4,6-2,2-1,8-3,8												0	3,2 0,60
M 7,8-2,6-1,4-6,6													0

MATRIZ DE DISTANCIAS DE CHEBYCHEV ENTRE 13 EJEMPLARES DE NEUROPTERIS

d_4 = CANBERRA	A	B	C	D	E	F	G	H	I	J	K	L	M
	3	3,2	2,3	2,6	3,7	3,9	4,9	6,6	6,9	5,7	4,8	4,6	7,8
	2,1	2,2	1,3	1,2	1,3	1,7	1,4	3,3	3,7	2,4	1,6	2,2	2,6
	1,2	1,4	1,4	1,2	1,7	1,2	0,7	2	2,1	1,1	1,4	1,8	1,4
	2,3	2,3	1,8	2	2,8	2,7	4,1	5,8	6,1	5	3,9	3,8	6,6
A 3-2,1-1,2-2,3	0	0,13 0,07	0,56 0,38	0,41 0,23	0,61 0,35	0,31 0,17	0,98 0,56	1,27 0,72	1,39 0,79	0,79 0,45	0,70 0,40	0,60 0,34	1,11 0,63
B 3,2-2,2-1,4-2,3		0	0,54 0,31	0,54 0,31	0,52 0,29	0,38 0,21	1,04 0,59	1,15 0,66	1,27 0,72	0,81 0,46	0,86 0,49	0,55 0,31	0,98 0,56
C 2,3-1,3-1,4-1,8			0	0,40 0,22	0,52 0,29	0,66 0,37	1,12 0,64	1,62 0,93	1,72 0,98	1,31 0,75	0,77 0,44	1,07 0,61	1,44 0,82
D 2,6-1,2-1,2-2				0	0,55 0,31	0,52 0,29	0,99 0,56	1,63 0,93	1,74 1	1,17 0,67	0,83 0,47	1,08 0,62	1,48 0,85
E 3,7-1,3-1,7-2,8					0	0,35 0,20	0,78 0,44	1,14 0,65	1,25 0,71	1 0,57	0,49 0,28	0,54 0,31	1,19 0,68
F 3,9-1,7-1,2-2,7						0	0,67 0,38	1,19 0,68	1,30 0,74	0,70 0,40	0,39 0,22	0,57 0,32	1,03 0,59
G 4,9-1,4-0,7-4,1							0	1,20 0,689	1,31 0,75	0,65 0,37	0,43 0,24	0,73 0,41	1,09 0,62
H 6,6-3,3-2-5,8								0	0,12 0,06	0,59 0,33	0,87 0,5	0,63 0,36	0,44 0,25
I 6,9-3,7-2,1-6,1									0	0,71 0,40	0,99 0,56	0,76 0,43	0,60 0,34
J 5,7-2,4-1,1-5										0	0,52 0,29	0,51 0,29	0,45 0,25
K 4,8-1,6-1,4-3,9											0	0,31 0,47	0,73 0,41
L 4,6-2,2-1,8-3,8												0	0,73 0,41
M 7,8-2,6-1,4-6,6													0

MATRIZ DE DISTANCIAS DE CANBERRA ENTRE 13 EJEMPLARES DE NEUROPTERIS

LA MATRIZ QUE HEMOS OBTENIDO PARA LA METRICA δ^* ENTRE LAS METRICAS d_1 , d_2, d_3 y d_4 ES :

δ^*	d_1	d_2	d_3	d_4
d_1	0	0,31 $L_2 - N_1$	0,2 $T_2 - L_3$	0,298 $S_3 - L_1$
d_2		0	0,291 $N_1 - N_2$	0,37 $H_1 - N_3$
d_3			0	$S_3 - L_2$ 0,38
d_4				0

METODO DEL ENLACE SENCILLO O DISTANCIA MINIMA

δ^*	d_1	d_2	d_3	d_4
d_1	0	0,31	0,20	0,298
d_2		0	0,291	0,37
d_3			0	0,38
d_4				0

d_1 Y d_3 SON FUSIONADOS PARA FORMAR UN GRUPO MUERTO QUE $\delta^*_{d_1, d_3} = 0,20$ ES EL MAS PEQUEÑO DE LOS ELEMENTOS DE LA MATRIZ.
LAS DISTANCIAS ENTRE ESTE GRUPO Y LOS DOS RESTANTES INDIVIDUOS $(d_2$ y $d_4)$ SE OBTIENEN A PARTIR DE δ^* DE LA SIGUIENTE FORMA :

$$\delta^*(d_1\ d_3)d_2 = \min\left\{\delta^* d_1\ d_2, \delta^* d_3\ d_2\right\} = \min\left\{0,31\ ,\ 0,291\right\} = 0,291$$

$$\delta^*(d_1\ d_3)d_4 = \min\left\{\delta^* d_1\ d_4, \delta^* d_3\ d_4\right\} = \min\left\{0,298, 0,38\right\} = 0,298$$

Y FORMAMOS UNA NUEVA MATRIZ DE DISTANCIAS δ_1 DANDO LAS DISTANCIAS INDIVIDUALES Y LAS DISTANCIAS GRUPO - INDIVIDUO

δ_1^*	$(d_1\ d_2)$	d_2	d_4	
$d_1\ d_3$	0	0,291	0,298	
d_2		0	0,37	
d_4			0	

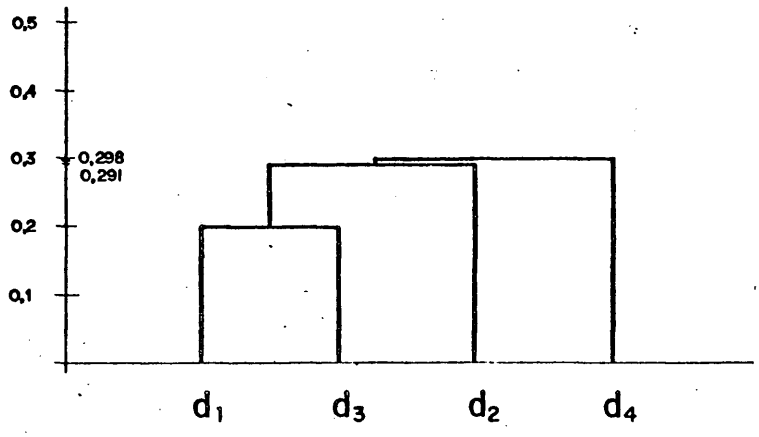
COMO EL ELEMENTO MAS PEQUEÑO DE LA MATRIZ ES 0,291 SE FORMA UN NUEVO GRUPO $[(d_1\ d_3)\ d_2]$

LA DISTANCIA δ_1 ENTRE EL NUEVO GRUPO $[(d_1, d_3)\ d_2]$ Y EL ELEMENTO d_4 ES:

$$\delta^* \left\{ \left[(d_1, d_3)\ d_2 \right], d_4 \right\} = \text{mín} \left\{ \left[(d_1\ d_3), d_4 \right], (d_2, d_4) \right\} = \text{mín} \left\{ 0{,}298\ ,0{,}37 \right\} = 0{,}298$$

Y YA ESTAN AGRUPADOS LOS 4 ELEMENTOS.

DENDROGRAMA :



METODO DE LA DISTANCIA PROMEDIO NO PONDERADO : GROUP AVERAGE

δ^*	d_1	d_2	d_3	d_4
d_1	0	0,31	0,20	0,298
d_2		0	0,291	0,37
d_3			0	0,38
d_4				0

PARTIMOS DEL NIVEL CERO Y HEMOS DE OBTENER EL NIVEL 1 A BASE DE AGRUPAR LOS DOS ELEMENTOS CUYA DISTANCIA SEA MINIMA, QUE EN ESTE CASO SON d_1 Y d_3

$\delta^* (d_1 , d_3) = 0,20$

LUEGO EL PRIMER GRUPO ES (d_1 , d_3)

$$\delta^* \left[(d_1 , d_3) , d_2 \right] = \frac{\delta^*(d_1 , d_2) + \delta^* (d_3 , d_2)}{2} = \frac{0,31 + 0,291}{2} = 0,30$$

$$\delta^* \left[(d_1 , d_3) , d_4 \right] = \frac{\delta^*(d_1 , d_4) + \delta^* (d_3 , d_4)}{2} = \frac{0,298 + 0,38}{2} = 0,339$$

LUEGO LA NUEVA MATRIZ ES :

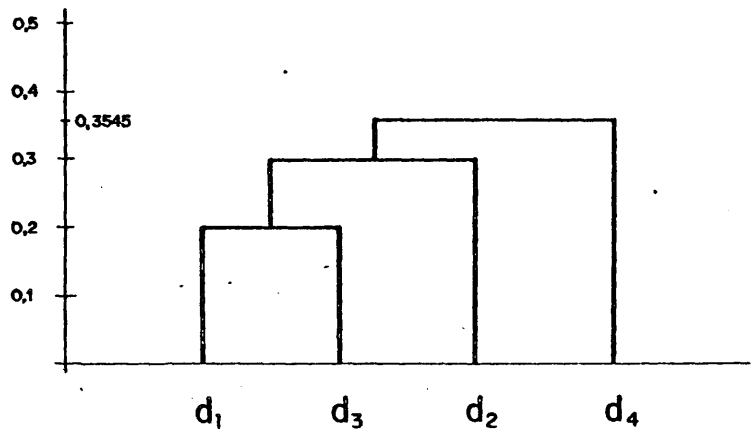
δ^*	$(d_1\ d_3)$	d_2	d_4
$(d_1\ d_3)$	0	0,30	0,339
d_2		0	0,37
d_4			0

COMO LOS ELEMENTOS QUE MENOR DISTANCIA TIENEN ENTRE SI SON $(d_1\ d_3)$ y d_2, 0,30, EL ELEMENTO d_2 SE FUSIONA CON EL GRUPO $(d_1\ d_3)$ [AL NIVEL DE DISTANCIA, 0,30]

Y, FINALMENTE, HEMOS DE HALLAR LA DISTANCIA ENTRE EL NUEVO GRUPO $[(d_1\ d_3), d_2]$ Y EL ELEMENTO d_4:

$$\delta^*\left\{\left[(d_1\ d_3)d_2\right], d_4\right\} = \frac{\delta_1^*\left[(d_1\ d_3)d_4\right] + \delta_1^*(d_2\ d_4)}{2} = \frac{0{,}339 + 0{,}37}{2} = 0{,}354$$

LUEGO EL DENDROGRAMA ES :



ENLACE COMPLETO (DISTANCIA MÁXIMA)

δ^*	d_1	d_2	d_3	d_4
d_1	0	0,31	0,20	0,298
d_2		0	0,291	0,37
d_3			0	0,38
d_4				0

COMENZAMOS, COMO EN EL METODO DEL ENLACE SENCILLO, FUSIONANDO LOS ELEMENTOS d_1 Y d_3 , $\delta^*(d\ ,d\ ) = 0,20$

LA DISTANCIA ENTRE ESTE GRUPO Y LOS DOS ELEMENTOS RESTANTES SE OBTIENE A PARTIR DE LA MATRIZ δ^* DE LA SIGUIENTE MANERA :

$$\delta^*\left[(d_1\ d_3), d_2\right] = \text{máx}\left[\delta^*(d_1\ , d_2), \delta^*(d_3\ , d_2)\right] = \text{máx}\left\{0,31\ ,\ 0,291\right\} = 0,31$$

$$\delta^*\left[(d_1\ d_3), d_4\right] = \text{máx}\left[\delta^*(d_1\ , d_4), \delta^*(d_3\ , d_4)\right] = \text{máx}\left\{0,298\ ,\ 0,38\right\} = 0,38$$

LUEGO LA NUEVA MATRIZ DE DISTANCIAS ES :

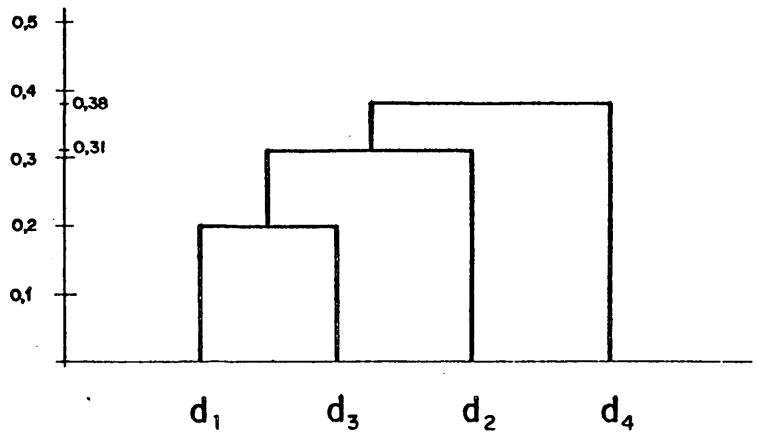
δ^*_1	$(d_1\ d_3)$	d_2	d_4
$(d_1\ d_3)$	0	0,31	0,38
d_2		0	0,37
d_4			0

COMO EL ELEMENTO MAS PEQUEÑO DE LA MATRIZ ES 0,31, SE FORMA, A ESE NIVEL, EL NUEVO GRUPO $\left[\left(d_1\ d_3\right), d_2\right]$
LA DISTANCIA ENTRE EL NUEVO GRUPO Y EL ELEMENTO d_4 ES LA SIGUIENTE :

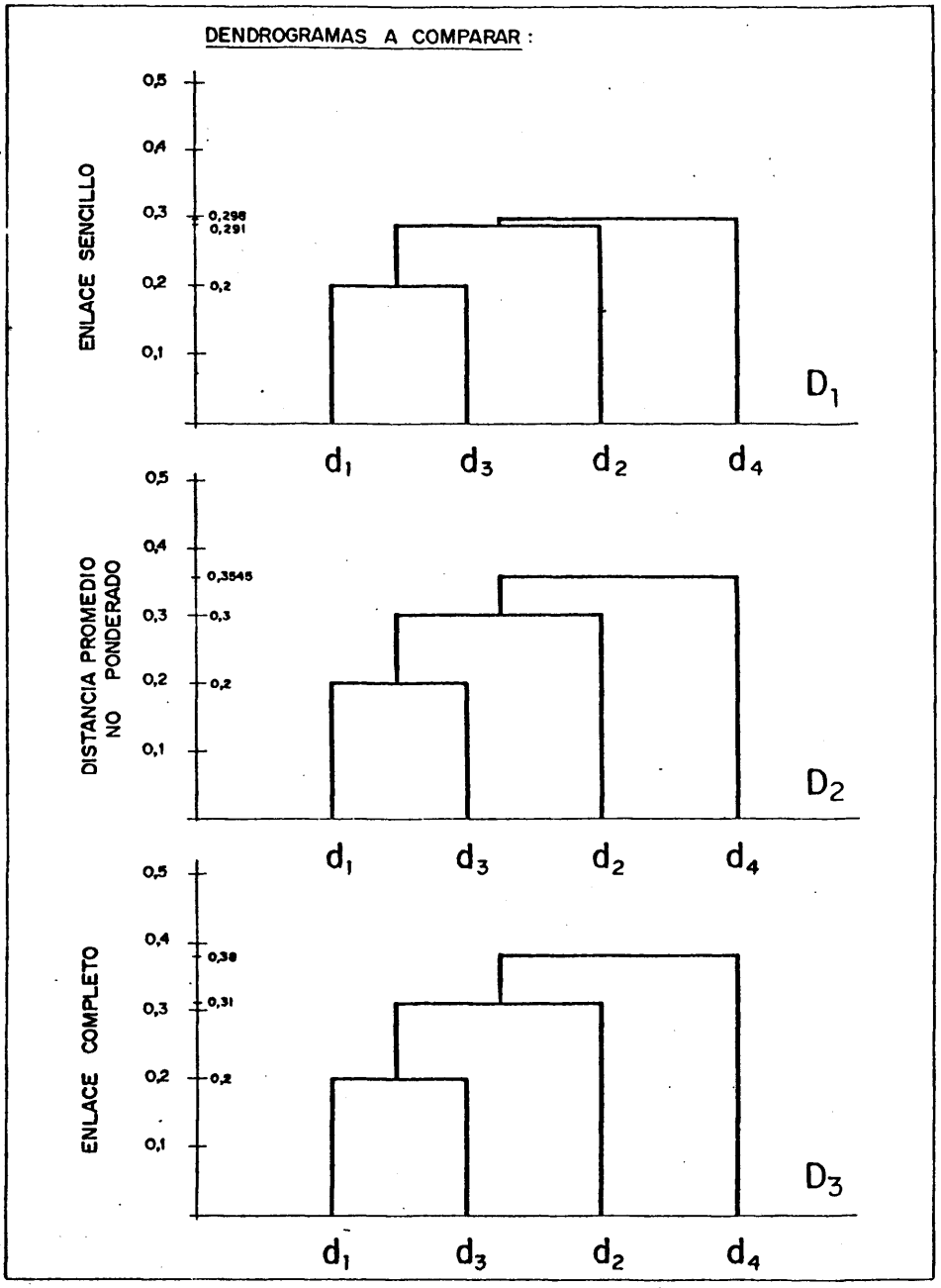
$$\delta^*\left\{\left[\left(d_1, d_3\right), d_2\right], d_4\right\} = \text{máx}\left\{\delta^*\left[\left(d_1, d_2\right), d_4\right], \delta^*\left[d_2, d_4\right]\right\} =$$

$$= \text{máx}\left\{0{,}38\ ,\ 0{,}37\right\} = 0{,}38$$

Y EL CORRESPONDIENTE DENDROGRAMA ES, FINALMENTE



DENDROGRAMAS A COMPARAR :



MATRICES DE VALORES COFENÉTICOS

$D_1 \longrightarrow$

	d_1	d_2	d_3	d_4
d_1	0	0,291	0,20	0,298
d_2	0,291	0	0,291	0,298
d_3	0,20	0,291	0	0,298
d_4	0,298	0,298	0,298	0

$D_2 \longrightarrow$

	d_1	d_2	d_3	d_4
d_1	0	0,30	0,20	0,3545
d_2	0,30	0	0,30	0,3545
d_3	0,20	0,30	0	0,3545
d_4	0,3545	0,3545	0,3545	0

$D_3 \longrightarrow$

	d_1	d_2	d_3	d_4
d_1	0	0,31	0,2	0,38
d_2	0,31	0	0,31	0,38
d_3	0,2	0,31	0	0,38
d_4	0,38	0,38	0,38	0

Al comparar los dendrogramas anteriores obtenemos los siguientes coeficientes de correlación:

$r_{12} = 0,993$

$r_{13} = 0,986$

$r_{23} = 0,998$

Y, finalmente, la matriz de coeficientes de correlación es:

r	D_1	D_2	D_3
D_1	\	0,993	0,986
D_2		\	0,998
D_3			\

Como se puede observar en la última matriz de coeficientes de correlación, los métodos de conglomeración más similares son los de la distancia promedio no ponderado y de la distancia máxima y los menos similares son los de la distancia mínima y el de la distancia máxima, lo que habíamos observado cuando hicimos tal comparación en el capítulo V .

RESUMEN

El proceso que hemos seguido para comparar las métricas EUCLIDEA, DE MANZANAS (CITY-BLOCK), CHEBYSHEV y CANBERRA, ha consistido en lo siguiente:

1.- Hemos definido un conjunto básico de OTUS.

2.- Hemos obtenido las matrices cuyos elementos son las distancias anteriores entre los OTUS del conjunto básico.

3.- A partir de dichas matrices y considerando como nuevos OTUS las métricas citadas, hemos obtenido, con la métrica δ^*, la correspondiente matriz de distancias.

4.- Con la última matriz de distancias y mediante los métodos jerárquicos del análisis de conglomerados, como son los de la distancia mínima, de la distancia promedio no ponderado y de la distancia máxima, hemos deducidos los correspondientes dendrogramas, cada uno de los cuales ya nos permite estudiar la similitud o disimilitud entre las métricas EUCLIDEA, DE MANZANAS, CHEBYSHEV y CANBERRA.

5.- Utilizamos el coeficiente de correlación cofenético para comparar las matrices de valores cofenéticos obtenidas a partir de los dendrogramas anteriores. Los valores del coeficiente de correlación cofenético nos permiten estudiar la similitud entre los dendrogramas anteriores y, por tanto, de los métodos de conglomeración.

BIBLIOGRAFIA

1.- AMON, J. (1980): "Estadística para psicólogos".
Pirámide.

2.- ANDERBERG, M.R. (1973): "Cluster analysis for applications".
Academic Press.

3.- ANDERSON, T.W. (1958): "Introduction to multivariate statistical analysis".
J. Wiley.

4.- ANDREWS, H.C. (1972): "Introduction to mathematical techniques in pattern recognition".
J. Wiley

5.- AZORIN POCH, F. (1962): "Notas sobre taxonomía y estadística".
Trabajos de Estadística e I.O., XXII.

6.- AZORIN POCH, F. (1972): "Curso de muestreo y aplicaciones".
Aguilar.

7.- AZORIN POCH, F. (1976): "Estadística y taxonomía matemática".
Estadística Española, I.N.E., 70 y 71.

8.- AZORIN POCH, F. (1972): "Estudio sobre la clasificación económica de los paises de America Latina".
Boletín Económico de América Latina. Naciones Unidas.
XVII.

9.- AZORIN POCH, F. (1979): "Conjuntos borrosos a la Estadística".
I.N.E.

10.- BAILEY, N. (1967): "The mathematical approach to Biology and Medicine".
J. Wiley.

11.- BALL, G.H. (1967): "A clustering technique for summarizing - multivariate data".
Behavioral Science, 12, 2.

12.- BALL, G.H. (1970): "Classification analysis"
Stanford Research Institute.

13.- BATCHELOR, B.G. (1974): "Practical approach to pattern classification".
Plemun Press.

14.- BENZECRI, J.P. (1976): "L'Analyse des donnees:T. II La Taxinomie".
Dunod.

15.- BENZECRI, J.P. (1976): "L'Analyse des donnes. T. 2: Correspondences".
Dunod.

16.- BERZTISS, A.T. (1975): "Data structures".
Academic Press.

17.- BLACKWELDER, R.E. (1967): "Taxonomy".
J. Wiley

18.- BOYCE, A.J. (1969): "Mapping diversity: A comparative study of some numerical methods".
"Numerical Taxonomy" COLE Ed. Academic Press.

19.- CACOULLOS, T. (1973): Discriminant analysis and applications".
Academic Press.

20.- CACOULLOS, T. (1973): "Distance, discrimination and error"
"Discriminant analysis" CACOULLOS, Ed. Academic Press.

21.- CARNAP, R.; MORGENSTERN, O. y WIENER, N (1974): "Matemáticas
en las ciencias del comportamiento".
Alianza Universidad.

22.- COLE, A. J. (1969): "Numerical Taxonomy".
Academic Press.

23.- CORMACK, R.M. (1971): "A review of classification".
Journal of the Royal Statistical Society.

24.- CRAMER, E.M. (1979): "Some symmetric, invariant measures of
multivariate association".
Psychometrica, V. 44, Nº 1.

25.- CRAMER, H. (1968): "Métodos matemáticos de Estadística".
Aguilar.

26.- CRAWFORD, R.M.M. (1969): The use of graphical methods in
classification".
"Numerical Taxonomy" COLE Ed. Academic Press.

27.- CROWSON, R.A. (1971): "Classification and Biology".
Heinemann Educational Books.

28.- CUADRAS, C.M. (1981): "Métodos de análisis multivariante". Editorial Universitaria de Barcelona.

29.- CHALON, E. (1975): "Reconocimiento y clasificación de patrones". Trabajos de Estadística e I.O. XXVI, pp. 93-135

30.- CHEN, C.H. (1973): "Statistical pattern recognition". Hayden.

31.- DAGET, J. (1976): "Les modeles mathematiques en Ecologie". Masson.

32.- D'ANDRADE, R.G. (1978): "U-Statistic hierarchical clustering". Psychometrika, V. 43, Nº 1.

33.- DAVIS, J.C. (1973): "Statistics and data analysis in Geology". J. Wiley.

34.- DUDA y HART (1973): "Pattern classification and scene analysis". J. Wiley.

35.- DURAN, B.S. y ODELL, P.L. (1970): "Cluster analysis" Board.

36.- EDWARDS, A.W. y CAVALLI-SFORZA, L.L. (1965): "A method for - cluster analysis". Biometrics, Junio.

37.- ELANDT-JOHNSON, R. (1971): "Probability models and statistical methods in Genetics".
John Wiley.

38.- ESCUDERO, L. (1968): "La estadística a la empresa".
I.E.E. (Madrid).

39.- ESCUDERO, L.(1974): "Estratificación tipológica: Nuevas posibilidades de la técnica Clustering".
U.A.M. - I.B.M. (Madrid).

40.- ESCUDERO, L. (1975): "Nuevos avances en las técnicas Clustering".
U.A.M. - I.B.M. (Madrid).

41.- ESCUDERO, L.F.(1977): "Reconocimiento de patrones".
Paraninfo.

42.- EVERITT, B. (1974): "Cluster analysis".
Heineman Educational Books.

43.- FELLER, W. (1958): "An introduction to probability theory and its applications".
J. Wiley.

44.- FISHER, I. (1967): "The making of index numbers".
Augusts M. Kelley.

45.- FISHER, W.D. (1958): "On grouping for maximun homogeneity".
J A S A. Diciembre.

46.- FISHER, W.D. (1969): "Clustering and aggregation in Economics" The Johns Hopkins Press.

47.- FRANZ, W. (1968): "Topología".
Selecciones Científicas.

48.- FUKUNAGA, K. (1972): "Introduction to statistical pattern recognition"
Academic Press.

49.- GARRIDO MARECA; J. (1976): "Taxonomía Matemática y Filosofía de las Formas de la Naturaleza".
R.A. de C.E.F. y N.

50.- GAYLORD, G. (1961): "Principles of animal taxonomy".
Columbia University Press.

51.- GENGERELLY, J.A. (1963): "A method for detecting subgroups in a population and sperifying their membership".
Journal of Phychology, 5

52.- GOOD, I.J. (1977): "The Botryology of Botryology in Classification and Clustering".
VAN RYZIN Ed. Academic Press.

53.- GOODMAN, L. y KRUSKAL, W.H. (1959): "Measures of association for cross classifications".
J A S A, Marzo.

54.- GORDON, A.D. (1981): "Classification".
Chapman and Hall.

55.- GORONZY, F. (1969): "A numerical taxonomy of business enterprises"
"Numerical Taxonomy". COLE Ed. Academic Press.

56.- GOWER, J.C. (1967): "A comparison of some methods of Cluster Analysis".
Biometrics, Diciembre.

57.- GRASSELI, A. (1969): "Automatic interpretation and classification of images".
Academic Press.

58.- GREGG, J. (1954): "The language of Taxonomy".
Columbia University Press.

59.- GUPTA, S.D. (1973): "Theories and methods in classification: A review".
"Discriminant Analysis". CACOULLOS, Ed.

60.- HALL, A.V. (1969): "Group forming and discrimination with homogeneity functions".
"Numerical Taxonomy". COLE Ed. Academic Press.

61.- HAND, D.J. (1981): "Discrimination and classification".
J. Wiley.

62.- HARBISON, F.; MARUHNIC, J. y RESNICH, J.R. (1970): "Quantita-

tive analysis of modernization and development".
Princenton University.

63.- HARRISON, P.J. (1968): "A method of Cluster Analysis and some applications".
Applied Statistics, 17

64.- HARTIGAN, J.A. (1975): "Clustering Algorithms".
J. Wiley.

65.- HARTIGAN, J.A. (1977): "Distributions problems in Clustering".
"Classification and Clustering". VAN RYZIN Ed. Academic Press.

66.- HEYWOOD, V.H. (1968): "Taxonomía vegetal".
Alhambra.

67.- HUBERT, L.J. (1978): "Evaluating the conformity of sociometric measurements".
Psychometrika, V. 43, nº 1.

68.- IVIMEY-COOK, R.B. (1969): "The phenetic relationsphips between species of Ononis".
"Numerical Taxonomy" COLE Ed. Academic Press.

69.- JACKSON, D.M. (1969): "Comparison of classifications".
"Numerical Taxonomy". COLE Ed. Academic Press.

70.- JARDINE, N. y SIBSON, R. (1971): "Mathematical Taxonomy".
J. Wiley.

71.- JAREÑO GARCIA, D. (1979): "Evaluación de sistemas taxonómicos basándonos en caracteres cualitativos".
Universidad Autónoma de Madrid.

72.- JOHNSON, S.C. (1967): "Hierarchical Clustering Schemes".
Phychometrika, 32.

73.- HALMOS, P.R. (1950): "Measure Theory".
Van Nostrand.

74.- HARRISON, P.J.: "A method of Cluster Analysis and some applications".
Applied Statistic.

75.- HOME OFFICE (1965): "Some statistical and other numerical techniques for classifying individuals".
Her Majesty's Stationary Office

76.- KAUFMAN, A. (1973): "Introduction a la théorie des Sous-Ensembles Flous".
Masson.

77.- KELLEY, J.L. (1955): "Topología General".
Eudeba.

78.- KENDALL, M.G. (1973): "The basic problems of Cluster Analysis".
"Discriminant Analysis". CACOULLOS Ed.

79.- KING, B. (1967): "Step-Wise Clustering Procedures".
A.S. Association Journal . Marzo.

80.- KOCH, G.S. y LINK, R.F. (1970): "Statistical Analysis of Geological data (I y II)".
J. Wiley.

81.- KOLMOGOROV, A.N. y FOMIN, S.V. (1975): "Elementos de la teoría de funciones y del análisis funcional".
Mir.

82.- KRUSKAL, J.B. (1964): "Nonmetric multidimensional scaling: A numerical method".
Psychometrika, 29

83.- KRUSKAL, J. (1977): "The relationship between multidimensional scaling and clustering".
"Classification and Clustering". VAN RYZIN Ed. Academic Press.

84.- LANCE, G.N. y WILLIAMS, W.T. (1967): "A general theory for classificatory sorting strategies. I. Hierarchical Systems.
Computer Journal, 9

85.- LANCE, G.N. y WILLIAMS, W.T. (1967): "A general theory for classificatory sorting strategies, II. Clustering Systems".
Computer Journal, 10

86.- LERMAN, I.C. (1969): "On two criteria of classification".
"Numerical Taxonomy". COLE Ed. Academic Press.

87.- LERMAN, I.C. (1970): "Les bases de la classification automatique".
Gauthiers-Villars.

88.- LOEVE, M. (1976): "Teoría de la probabilidad".
Tecnos.

89.- MATULA, D.W. (1977): "Graph theoretic techniques for Cluster Analysis Algorithms".
"Classification and Clustering". VAN RYZIN Ed. Academic Press.

90.- MOLINA CANO, J.L. (1977): "Introducción a la Taxonomía Numérica".
Monografía de la E.T.S.I. Agronomos de Madrid.

91.- MORRISON, D.G. (1967): "Measurement problems in Cluster Analysis".
Management Science, V. 13, nº 12.

92.- MUNROE, M.E. (1953): "Introduction to measure and integration".
Addison-Wesley.

93.- ORLOCI, L. (1969): "Information theory models for hierarchic and nonhierarchic classifications"
"Numerical Taxonomy". COLE Ed. Academic Press

94.- PEREZ BEATO, M. (1979): "Bioestadística".
Madrid.

95.- PEREZ CAMACHO, F. (1973): "Aplicación de la Taxonomía Numérica a variedades cultivadas de olivo".
Fundación Juan March.

96.- PFLAUMANN, E. y UNGER, H. (1968): "Análisis funcional".
Alhambra.

97.- RAO, R.C. (1977): "Cluster Analysis applied to a study of race mixture in human populations".
"Classification and Clustering". VAN RYZIN Ed. Academic Press.

98.- RIOS, S. (1967): "Métodos estadísticos".
Ed. del Castillo (Madrid).

99.- RIOS, S. (1976): "Análisis de decisiones".
Ediciones I.C.E. (Madrid).

00.- RIVAS MARTINEZ, S. (1975): "Perspectivas sobre taxonomía vegetal".
R. A. de Farmacia.

01.- ROHLF, F.J. (1975): "A new approach to the computation of the Jardine-Sibson B_k Clusters".
Computer Journal, 18.

02.- ROSS, G.J.S. (1969): "Classification techniques for large -- sets of data".
"Numerical Taxonomy". COLE Ed. Academic Press

103.- ROUX, M (1969): "An Algorithm to construct a particular kind of hierarchy"
"Numerical Taxonomy". COLE Ed. Academic Press.

104.- SANCHEZ GARCIA, M.: "Reconocimiento de formas".

105.- SANCHEZ GARCIA, M.: "Conglomerados y sus aplicaciones".
Coloquio Internacional de Estadística e I.O. Madrid.

106.- SCHREIDER, J.A. (1975): "Equality, resemblance and order".
Mir Publishers.

107.- SEBESTYEN, G. (1962): "Decisión-making processes in pattern recognition".
Mac Millan.

108.- SHEPARD, R. KIMBALL ROMNEY, A. y BETH NERLOVE, S. (1972):
"Multidimensional scaling. Vol. I y II".
Seminar Press.

109.- SIBSON, R. (1972): "Order invariant methods for data analysis".
Royal Statistical Society. Mayo.

110.- SNEATH, P.H.A. (1967): "Some statistical problems in numerical taxonomy".
The Statistician, 17

111.- SNEATH, P.H.A. (1969): "Evaluation of Clustering Methods".
"Numerical Taxonomy". COLE Ed. Academic Press.

112.- SNEATH y SOKAL, (1973): "Numerical Taxonomy".
W.H. Freeman.

113.- SOKAL, R.R. (1977): "Clustering and classification: Background and current directions".
"Classification". VAN RYZIN Ed. Academic Press.

114.- SOKAL, R. y ROHLF, F.J. (1969): "Biometría".
H. Blume Ediciones.

115.- SOKAL, R. y SNEATH, P.H. (1963): "Principles of numerical Taxonomy".
W.H. Freeman.

116.- SOLOMON, H. (1977): "Data dependent Clustering techniques".
"Classification and Clustering". VAN RYZIN. Academic Press.

117.- THEIL, H. (1967): "Economics and information theory".
Rand Mc Nally.

118.- THEIL, H. (1975): "Theory and measurement of consumer demand (V. I)".
North-Holland.

119.- TRILLAS RUIZ, E. (1973): "Sobre distancias estadísticas".
Tesis Doctoral, Barcelona.

120.- TRYON, R.C. y BAYLEY, D. (1970): "Cluster Analysis".
Mc Graw Hill.

121.- TURNER, J.C. (1970): "Matematica moderna aplicada: Probabilidades, estadística e I'operativa".
Alianza Editorial.

122.- VALLE SANCHEZ, A. (1963): "La integral de Lebesgue. Teoría elemental de las distribuciones de Schwartz".
C.S.I.C.

123.- VAN RYZIN, J. (1977): "Classification and Clustering".
Academic Press.

124.- VEGAS PEREZ, A. (1980): "Estadística: Aplicaciones econométricas y actuariales".
Pirámide.

125.- WATERMAN, T. y MOROWITZ, H. (1965): "Theoretical and mathematical Biology".
Blaisdell Publishing Company.

126.- WILLIAMSON, M. (1972): "The analysis of biological populations".
Edward Arnold.

127.- WILLMOTT, A.J. y GRIMSHAW, P.N. (1969): "Cluster analysis in Social Geography".
"Numerical Taxonomy". COLE Ed. Academic Press.

128.- WISHART, D. (1969): "Mode analysis: A generalisation of near rest neighbour which reduces chaining effects".
"Numerical Taxonomy". COLE Ed. Academic Press.

29.- YOUNG, T. y CALBERT, T.W. (1974): "Classification, estimation and pattern recognition".
Elsevier.

130.- ZURANO HERNANDEZ, M. A. (1979): "Clasificaciones aceptables basadas en caracteres cuantitativos".
Universidad Autónoma de Madrid.

INDICE DE AUTORES

INDICE ALFABETICO